Aalto University
School of Science
Master's Programme in Mathematics and Operations Research

Joonas Laihanen

Distribution-based Subpopulation Framework and on Its Applications in the Aviation Industry

Master's Thesis
Espoo, June 7, 2019

Supervisor: Asst. Prof. Pauliina Ilmonen, Aalto University
Advisor: M.Sc. (Tech.) Johanna Tikanmäki, Finnair

Aalto University
School of Science
Master's Programme in Mathematics and Operations Research

ABSTRACT OF MASTER'S THESIS

Author: Joonas Laihanen

Title:
Distribution-based Subpopulation Framework and on Its Applications in the Aviation Industry

Date: June 7, 2019 **Pages:** 84

Major: Systems and Operations Research **Code:** SCI3055

Supervisor: Asst. Prof. Pauliina Ilmonen

Advisor: M.Sc. (Tech.) Johanna Tikanmäki

Many real-life data sets have subpopulations in them, which have similarly shaped probability distributions for their variables. We aim to build a framework to understand the data better by examining this subpopulations structure, finding the common patterns in shapes between the subpopulations and utilizing the found distributions for applications. The framework presented in this thesis is an illustrative example how this can be done in practice.

The framework is a process that plainly has four steps: the data is clustered to subpopulations, the density is estimated for each subpopulation, the common shape for the subpopulations is determined and improved estimates for the subpopulation densities are calculated. Each step provides understanding about the patterns in the data. The framework applies variety of methods and concepts from the fields of data mining and statistics, such as statistical moments and L-moments, moment diagrams, regression, model-based and hierarchical clustering, parametric families, mixture models, kernel density estimation, location-scale families and introduces a few novel ideas, definitions and algorithmic methods in order to do each step of the framework.

The framework is demonstrated by applying it to case examples from the aviation industry. The main case is about understanding the passenger weight data: how the groups with a low number of samples are truly distributed, can the sample size be reduced without accuracy loss, what kind of standard weights should be used and how should the passengers be segmented and can the extreme values be estimated. We also consider how the weather affects the flight delays and further usage of the framework.

Overall, the framework works well in the aviation cases: meaningful subpopulations with similar distribution structure are found, the improvements for the estimates of the distribution shapes are significant when using the found common shape and the practical goals in the aforementioned passenger weight case are met. The framework seems very prominent for understanding multivariate data sets in nature and business using the distributional subpopulation structure within the data associated to them.

Keywords: probability distribution, moments, clustering, shape identification, subpopulation structure, passenger weights, aviation

Language: English

Aalto-yliopisto
Perustieteiden korkeakoulu
Matematiikan ja operaatiotutkimuksen maisteriohjelma

DIPLOMITYÖN TIIVISTELMÄ

Tekijä: Joonas Laihanen

Työn nimi:
Jakaumapohjainen alipopulaatioviitekehys ja sen sovelluksista lentoliiketoiminnassa

Päiväys:	7. kesäkuuta 2019	**Sivumäärä:**	84
Pääaine:	Systeemi- ja operaatiotutkimus	**Koodi:**	SCI3055
Valvoja:	Asst. Prof. Pauliina Ilmonen		
Ohjaaja:	DI Johanna Tikanmäki		

Monissa tosielämän datajoukoissa esiintyy alipopulaatioita, joilla on samanmuotoisia todennäköisyysjakaumia muuttujilleen. Pyrimme rakentamaan viitekehyksen ymmärtääksemme dataa paremmin tutkimalla näitä alipopulaatioita, etsimällä säännönmukaisuuksia niissä ja käyttämällä löydettyjä jakaumia sovelluksissa. Tässä diplomityössä esitettävä viitekehys on havainnollistava esimerkki siitä, kuinka tällainen kokonaisuus voidaan toteuttaa käytännössä.

Viitekehyksessä on yksinkertaistetusti neljä vaihetta: datan klusterointi alipopulaatioihin, jakauman arvioiminen jokaiselle alipopulaatiolle, jakaumien yhteisen muodon tunnistaminen ja parannettujen arvioiden määrittäminen alipopulaatiojakaumille. Jokainen vaihe auttaa ymmärtämään datan erinäisiä säännönmukaisuuksia. Viitekehys käyttää useita metodeja ja käsitteitä tiedonlouhinnan ja tilastotieteen aloilta, kuten tilastollisia momentteja, L-momentteja, momenttikaaviota, regressiota, mallipohjaista ja hierarkista klusterointia, parametrisiä jakaumaperheitä, yhdistelmämalleja, ydinestimointia, lokaatio-skaalausjakaumaperheitä ja esittelee vereviä ajatuksia, määritelmiä ja algoritmeja viitekehyksen vaiheiden toteutukseen.

Viitekehystä demonstroidaan soveltamalla sitä tapausesimerkkeihin lentoliiketoiminnasta. Pääesimerkkinä tutkitaan, kuinka matkustajapainodataa voidaan ymmärtää: kuinka pienen otoskoon ryhmät ovat todella jakautuneet, voiko otoskokoa pienentää tarkkuuden kärsimättä, mitä standardipainoja tulisi käyttää ja miten matkustajat tulisi segmentoida ja voidaanko ääriarvoja arvioida. Pohdimme myös sään vaikutusta lentojen myöhästymisiin ja viitekehyksen sovellettavuutta laajemmin.

Yleisesti ottaen viitekehys toimii hyvin lentoliiketoiminnan ongelmissa: mielekkäitä alipopulaatioita samankaltaisilla jakaumamuodoilla löydetään, jakaumamuotoja saadaan arvioitua merkittävästi paremmin löytämällä yhteinen jakaumamuoto ja tavoite ymmärtää matkustajapainodataa toteutuu hyvin. Viitekehys vaikuttaa lupaavalta datajoukkojen ymmärtämiseen luonnossa ja liiketoiminnassa tunnistamalla alipopulaatiorakenteita sovelluskohteiden datoissa.

Asiasanat:	todennäköisyysjakauma, momentit, klusterointi, jakaumamuodon tunnistus, alipopulaatiorakenne, matkustajapainot, lentoliikenne
Kieli:	Englanti

Acknowledgements

This thesis has been done in collaboration with Finnair. I thank people at Finnair for offering a chance to do this thesis project along with related data analysis tasks. Finnair gave me a lot of freedom to determine the contents of the thesis, but also offered essential support, when trying to setting up interesting questions to resolve and offered the necessary data resources. To mention a few key people, I would like to express my gratitude for my manager Hannu Hakkarainen, thesis advisor Johanna Tikanmäki, everyone who contributed to the passenger weight project and the field work of collecting the weight samples like Juan Méndez Jiménez and many others. The thesis was also supposed to include the weather-related delay case in more detail, but due to the problem being too extensive for the scope of the thesis, it is diminished in this written work. However, I thank people who showed interest in the weather and delay project including Pekka Lappi from Finnair, Juha Fieandt from SureWx and Kari Österberg from Finnish Meteorological Institute.

From the academia, I foremost thank my supervisor Pauliina Ilmonen, who was always there to meet and discuss about the thesis project or other aspects of my studies. I also thank everyone else in Aalto, the department of mathematics and the systems analysis lab, who provided insightful comments to my questions about this work and to everything else too.

Lastly, I would like to sincerely thank my folks, not necessary for contributing anything to this very thesis, but for the other efforts in life.

Helsinki, June 7, 2019

Joonas Laihanen

Contents

Chapter 1

Introduction to the Framework

Understanding patterns and predicting quantities of interest using large data sets have become increasingly prevalent in our time. The field of data mining, or more descriptibly known as knowledge discovery from data [17], along with statistics and machine learning, try to answer these kind of questions.

In this thesis we introduce a framework consisting of data sources, statistical and data mining methods and an operational business problem of interest. The framework is an illustrative example approach of how one can utilize the underlying subpopulation structure of the data and probability density distribution modeling. The construction of the framework is originally motivated by practical problems in the aviation industry.

This introduction chapter provides the background for the thesis: we explain our motivations and goals further, what has been done previously to tackle theses issues and the structure of the thesis by giving an overview of the framework.

1.1 Context and the Framework Overview

Many businesses have large amounts of data to improve their decision making and face the need of using advanced data mining methods [36], on top of basic data analysis, which may only superficially explain the patterns in the

data, and miss the true causes and interconnections within the data, and as so, within their businesses. For example, customer segmentation, demand prediction, risk evaluation of rare events and many more can be understood statistically using data mining methods [31]. A data set consisting of a collection of multivariate samples is a common type of a data set, but in order to understand general patterns in it, one sample alone provides little information, and we need to comprehend the big picture by, for instance, comparing variable correlations or clustering the samples into groups. The latter enables estimating empirical distributions, making visualizations and understanding the characteristics of populations. Many business experts, as an example McAfee [28], state that big data and data-driven decision making are key issues in improving company's performance and competitive edge, but also acknowledgement that to use the power of big data a greater vision and human insights are required.

In general, statistical assumptions are needed when inferring conclusions from real-life data. A common practice is making the normality assumption, that is assuming that the data is distributed as Gaussian. This Gaussian framework has a solid theoretical foundation that allows easy computations of confidence intervals and many other benefits, and it is widely used especially in research of medicine, psychology and social sciences. However, it fails to catch the characteristics of unsymmetrically distributed observations and fails to model a few conditions such as that all values must be non-negative for certain type of variables. [27] One of our main motivations is to challenge this limited Gaussian way of understanding the structure of distributions in the data. We assume that the data has some similarly shaped subpopulations and use this to catch the true characteristics of the subpopulations. This assumption is natural as it is known that many such structures exist in nature, for instance human heights, personal wealth and ages each seem to follow a certain type of distribution family among their subpopulations, such as the nationality in the world. We call this the distribution-based subpopulation framework (abbr. DS-framework). We attempt to avoid of making unnecessary assumptions and make our framework as general as possible, but there are a few exceptions, including that in one of the methods we introduce, we do make the location-scale assumption which assumes that the subpopulations are from the same location-scale family.

The DS-framework consist of four main stages: clustering the data to subpopulations, making the probability density estimations for the subpopulations, identifying the common shape for the subpopulations and applying the found common shape to improve the density estimations. The framework is

presented in Figure 1.1.

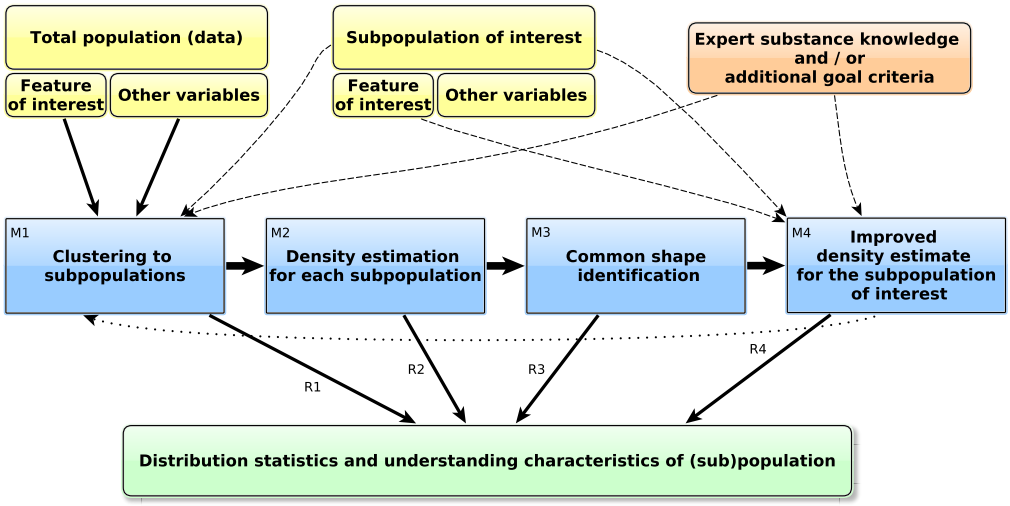


Figure 1.1: A chart of the distribution-based subpopulation framework used in this thesis.

In practice, we assume here that we have a data set that consist of observations, which have multiple attributes. To understand certain aspect of the data we choose a feature of interest for which the distributions are built. The feature of interest can be a variable in the data or anything that can be calculated using the data, and can be connected to each observation. Other variables in the data are used mainly for the clustering purposes. Also we might have a subpopulation of interest, a particular group, for which we want to make the distribution estimations, instead of just for all subpopulations that results from the clustering. In this case, the subpopulation of interest should also be used as a criterion in the clustering. In many cases, it is advisable to do the clustering completely based on the subpopulation of interest, if it exist in the framework. In addition, we might have some external domain knowledge, which can be used in clustering or we might have external criteria on how we should frame the problem, such as the need that the subpopulation means differ as much as possible. So overall, our framework is essentially a process that goes back and forth for the best results. It can be used as an algorithm that has four stages, but for the best results and to truly understand the feature of interest, density estimation should also affect the clustering. Each stage provides understanding about the data in the form of providing different probability distributions and clustering results.

As a comparison, there many other approaches, which attempt to under-

stand the subpopulation of interest by borrowing strength from other subpopulations, data or domain knowledge in general. For example, the area of small area estimation, see research by Rao [37] and [13], or newer design and model-based approaches [35], which examine how statistic estimates can be made for subpopulations with a low number of samples. On top of this, there are other many approaches, that also may refer to themselves with the term framework, which solve similar problems than our framework, with and without the Gaussian modeling, see for instance, [11], [2],[24],[18] [38] and [7]. However, our approach includes the usage of distribution shape identification, an idea that has been used in other type of settings like [33] and [45], and the framework is built to be dynamic for different data sets and operational interests.

To sum up, our main goal is to **understand the data using the phenomenon of subpopulations having similar distributional shapes within the real life data**. Therefore the framework provides many insights about the feature of interest and the data, for instance:

- Meaningful segmentation of the data
- Event estimation, for example, to estimate variance-at-risk
- Better understanding of phenomena in the world and subpopulations within them
- Improved event estimation, especially for rare events

1.2 Structure of the Thesis

This thesis starts with an overview of the DS-framework in this chapter. The following chapters go through the four blocks of the framework. Chapter 2 considers the first two blocks of the framework. It starts by introducing basics concept for the rest of thesis, then explains the distribution fitting and clustering methods relevant for our research interests. Chapter 3 presents how the common shape can be identified and the improved estimate for the subpopulation of interest can be constructed. Many of the examples in these chapter use parts of the passenger weight data set that is overviewed in 4.2.

Chapters 4 and 5 provide case examples that demonstrate the usage of the framework. The cases are from the aviation industry and the problems presented in these cases are partly the initial motivation why the framework has been structured as it is. Chapter 4 considers implications that passenger weights and newly got measurements of those have for the airline company of interest. Chapter 5 examines shortly further applications, such as the relation that the weather has to the flight departure delays, and concludes the thesis.

The methods presented in this thesis are meant to be instructive examples, and many more alternative constructions for these kind of frameworks can be developed and should be utilized, for instance, if the variable of interest would be discrete. Note that though all the theory presented will not be applied in our case examples, it should be rather effortless for an educated reader to see how different alternative methods would be used in practice.

Chapter 2

On Distribution Fitting and Clustering to Subpopulations

We start by introducing relevant statistics for probability density distributions, which are the foundation for the concept of subpopulations having a similar shape and for other parts of the theory. Then we construct the framework in more detail by defining the first two blocks, that is clustering the population to subpopulations and fitting a distribution to samples of each subpopulation.

2.1 Basic Concepts to Build the Framework

A distribution can be understood mechanically by its numerical statistics. Many of them relate to the shape of the distribution, but the question whether two distributions are similarly shaped is challenging.

2.1.1 Characteristics of Distributions

Probability distributions have a wide range of different measures that describe their location, variation and shape. Moments of various degrees are related to these quantities. A few of basic moment concepts are:

Definition 1. *(n^{th} moment of a random variable): The n^{th} moment of a random variable with PDF f is*

$$\mu_n = \mathrm{E}\left[X^n\right] = \int_{-\infty}^{\infty} x^n \, f(x) \, \mathrm{d}x, \tag{2.1}$$

where E *is the expected value operator.*

Definition 2. *(n^{th} central moment of a random variable): The n^{th} central moment of a random variable with a PDF f is*

$$\mu_n^{'} = \mathrm{E}\left[(X - \mathrm{E}[X])^n\right] = \int_{-\infty}^{+\infty} (x-\mu)^n f(x) \, \mathrm{d}x. \tag{2.2}$$

Definition 3. *(n^{th} normalized moment of a random variable): The n^{th} normalized moment of a random variable is*

$$\mu_n^* = \frac{\mu_n}{\sigma^n} = \frac{\mathrm{E}\left[(X-\mu)^n\right]}{\sigma^n}, \tag{2.3}$$

where σ is the standard deviation.

As an example, the first degree moment is the mean value, which describes where the data is located, the second degree tells about the variation and the second central moment is called the variance. Furthermore, the third normalized moment skewness is a measure of asymmetry that tells which of the tails of a distribution is heavier. The fourth normalized moment kurtosis tells how much of the probability is in the tails versus around the center. See Reference [41] for basic moment and distributions concepts.

When we do not know the underlying distribution, but only have samples from the distribution, we estimate the moments using sample moments. For instance, for the samples $X_1, ..., X_N$ the n^{th} sample moment can be estimated by

$$\hat{\mu}_n = \frac{1}{N} \sum_{i=1}^{N} X_i^n \tag{2.4}$$

These estimators can be defined in a variety of ways, different estimators having different properties, such as, distinct bias, consistency, efficiency and robustness properties. The basic theory of sampling distributions and estimators can be found, for example, in Book [44] and for the more advanced properties see [6], [41] and [25].

Other useful alternatives that are used in this thesis to understand the distributions include using different order statistics. Well-known examples of order statistics for continuous distributions are quantiles, such 2-quantile that is the median and 100-quantile that is the percentile. The $(1-\alpha)$-quantile of a random variable $-X$ is called the Value-at-Risk (VaR) of X at the level α. This concept is frequently applied in the context of risk analysis.

Here to find patterns within the distributions, we introduce useful statistics called L-moments, which are linear combinations of order statistics. [21] L-moments are more robust than the regular moments, require only a finite mean to exist and possibly reveal patterns that cannot be seen with the regular moments. These can be defined:

Definition 4. *(n^{th} L-moment of a random variable): The n^{th} L-moment of a random variable X is*

$$\lambda_n = n^{-1} \sum_{k=0}^{n-1} (-1)^k \binom{n-1}{k} \mathrm{E}X_{n-k:n}, \tag{2.5}$$

in which $X_{k:N}$ is the k^{th} smallest value in an independent sample of N samples from the distribution X.

Furthermore, similarly to regular moments we can consider the scaled versions:

Definition 5. *(n^{th} L-moment ratio of a random variable): The n^{th} L-moment ratio is*

$$\tau_n = \lambda_n / \lambda_2. \tag{2.6}$$

Moreover, we can quantify the similarity of two distributions by different metrics. A traditional method to compare distributions is to plot their quantiles against each other. This is called Q-Q plotting. For similar distributions the values of this plot should lie on the line $x = y$. The earth mover's distance (EMD) assign a single value for the distance between two distributions using the same premise as Q-Q plotting in the setting we are considering, but note that in a more general settings the EMD is defined more precisely and has been applied for applications, such as image retrieval [39] and graph comparision [45]. For two real-valued probability distributions $f_1(x)$ and $f_2(x)$ with the area of 1, the definition simplifies to:

Definition 6. *(EMD for two probability distributions): The simplified EMD distance metric between two distributions is*

$$\int_{x=-\infty}^{\infty} \left| \int_{y=-\infty}^{x} (f_1(y) - f_2(y))\, dy \right| dx. \tag{2.7}$$

On top of considering the whole interval it is possible to calculate this value for top or bottom quantiles only to access how good a fit is for the extreme values. This is essentially how we predict the extreme values in this study. More refined alternatives for understanding extreme values include using extreme value estimators, like Hill estimator [20] for heavy-tailed distributions, and other approaches from the extreme value theory.

2.1.2 Families of Distributions and Patterns in Them

Assuming an underlying distribution family is a common practice in statistics and applications, as many statistical parametric methods rely on this [5], as it not only enables an easy determination of the distribution by calculating the optimal parameters, but also because in many cases we have domain knowledge how the variables should be distributed in theory. Example of commonly used families of continuous distributions include, normal, beta, gamma, exponential and log-normal distribution families. Many families have been developed for specific purposes in mind, such as Gumbel distributions for the extreme value prediction and survey analysis [15].

The parameters in the parametrically defined families of distributions are usually either location, scale or shape parameters. For instance, the normal distribution is controlled by location parameter μ that also is the mean value, and scale parameter σ that is the deviation. It is common that location parameters are strongly connected to the first moment, scale parameters to the second, and shape for the higher moments. Location and scale are often rather effortlessly definable in the case of parametric families but the shape is more ambiguous. Intuitively we would like to have the possibility that the location and scale can be different, but other properties should not. If this is true, then the two distributions could be considered to be similarly shaped. For instance, normal distributions exhibit this kind of behavior, as they are defined by the location and shape parameter completely and their upper moments are determined by their location and scale. This provides two intuitive ways of defining the shape of a distribution, and as so, whether

two distribution are similarly shaped. Naively, using the first two moments we could propose:

Definition 7. *(Shape similarity with the first two moments): Functions in a family of distributions $\mathcal{F}$ are shaped similarly if for any $f_1, f_2 \in \mathcal{F}$, it holds that if $\mu_{1,f_1} = \mu_{1,f_2}$ and $\mu_{2,f_1} = \mu_{2,f_2}$ then $F_1 = F_2$.*

Definition 8. *(Location-scale family): Assuming that random variable X and Y have probability density functions $f_X(x)$ and $f_Y(y)$, such that it holds $f_Y(y) = \frac{1}{b} f_X\left(\frac{y-a}{b}\right)$ for some $a \in \mathbb{R}$ and $b \in \mathbb{R}_+$, then X and Y belong to the same location-scale family.*

Definition 9. *(Shape similarity in the location-scale family): The random variables X and Y are shaped similarly if they belong to the same location-scale family.*

Naturally, if all our distributions are clearly from the same location-scale family this is very compelling definition, but in many cases the shapes can be more complex and more complicated definitions are required. In our further approaches we suggest that the distributions might have some recognizable patterns in them that can be identified using the moments. We focus on usual moments, L-moments and simple relations calculated from them to find patterns, but a reader may imagine more sophisticated or alternative methods that could be based on quantiles or other well-known statistics.

For instance, in the family of normal distributions the central moments are $\mu_n^{'} = \sigma^{2n} * (2n-1)!!$ for even n, double exclamation mark meaning the semifactorial, and $\mu_n^{'} = 0$ for odd n. [46] Now, assuming that we do not know that the samples we observe are from normal distributions, we still might be able to see that the odd central moments are close to 0 and even recognize the more advanced pattern for the even moments, if we have enough samples and the number of distributions is also sufficiently large.

The moment patterns for two fixed degrees can be systemically observed in moment-ratio diagrams [43], although the selection of which moments, moment ratios or other statistics should be visualized is not trivial. Especially, L-moments seem to have very benign patterns to understand the common shape of a family and provide a procedure to select the distribution family according to the data. [34]

Table 2.1: Different L-moments and L-moment ratios [21]

L-moments of some common distributions†

Distribution	*F(x) or x(F)*	*L-moments*
Uniform	$x=\alpha+(\beta-\alpha)F$	$\lambda_1=\frac{1}{2}(\alpha+\beta),\ \lambda_2=\frac{1}{6}(\beta-\alpha),\ \tau_3=0,\ \tau_4=0$
Exponential	$x=\xi-\alpha\log(1-F)$	$\lambda_1=\xi+\alpha,\ \lambda_2=\frac{1}{2}\alpha,\ \tau_3=\frac{1}{3},\ \tau_4=\frac{1}{6}$
Gumbel	$x=\xi-\alpha\log(-\log F)$	$\lambda_1=\xi+\gamma\alpha,\ \lambda_2=\alpha\log 2,\ \tau_3=0.1699,\ \tau_4=0.1504$
Logistic	$x=\xi+\alpha\log\{F/(1-F)\}$	$\lambda_1=\xi,\ \lambda_2=\alpha,\ \tau_3=0,\ \tau_4=\frac{1}{6}$
Normal	$F=\Phi\left(\frac{x-\mu}{\sigma}\right)$	$\lambda_1=\mu,\ \lambda_2=\pi^{-1}\sigma,\ \tau_3=0,\ \tau_4=30\pi^{-1}\tan^{-1}\sqrt{2}-9=0.1226$
Generalized Pareto	$x=\xi+\alpha\{1-(1-F)^k\}/k$	$\lambda_1=\xi+\alpha/(1+k),\ \lambda_2=\alpha/(1+k)(2+k),$ $\tau_3=(1-k)/(3+k),\ \tau_4=(1-k)(2-k)/(3+k)(4+k)$
Generalized extreme value	$x=\xi+\alpha\{1-(-\log F)^k\}/k$	$\lambda_1=\xi+\alpha\{1-\Gamma(1+k)\}/k,\ \lambda_2=\alpha(1-2^{-k})\Gamma(1+k)/k,$ $\tau_3=2(1-3^{-k})/(1-2^{-k})-3,$ $\tau_4=(1-6.2^{-k}+10.3^{-k}-5.4^{-k})/(1-2^{-k})$
Generalized logistic	$x=\xi+\alpha[1-\{(1-F)/F\}^k]/k$	$\lambda_1=\xi+\alpha\{1-\Gamma(1+k)\,\Gamma(1-k)\}/k,\ \lambda_2=\alpha\Gamma(1+k)\,\Gamma(1-k),$ $\tau_3=-k,\ \tau_4=(1+5k^2)/6$
Log-normal	$F=\Phi\left(\frac{\log(x-\xi)-\mu}{\sigma}\right)$	$\lambda_1=\xi+\exp(\mu+\sigma^2/2),\ \lambda_2=\exp(\mu+\sigma^2/2)\,\mathrm{erf}(\sigma/2),$ $\tau_3=6\pi^{-1/2}\int_0^{\sigma/2}\mathrm{erf}(x/\sqrt{3})\exp(-x^2)\mathrm{d}x/\mathrm{erf}(\sigma/2)$
Gamma	$F=\beta^{-\alpha}\int_0^x t^{\alpha-1}\exp(-t/\beta)\mathrm{d}t/\Gamma(\alpha)$	$\lambda_1=\alpha\beta,\ \lambda_2=\pi^{-1/2}\beta\Gamma(\alpha+\frac{1}{2})/\Gamma(\alpha),\ \tau_3=6I_{1/3}(\alpha,2\alpha)-3$

†γ is Euler's constant; Φ is the standard normal distribution function; $I_x(p, q)$ is the incomplete beta function.

In Table 2.1, that is directly taken form the article by Hoskings [21], L-moment ratios are shown for a few well-known distribution families. There seems to be notable patterns, mainly that the third and forth ratios being constant, whereas the first and second ratios seems to depend on the parameters for many families. Observing this kind of behavior, when the ratios are constant or otherwise behave in predictable manner, such as being on a line, curve or on certain area, could be a way to understand, when the distributions are shaped similarly. Figure 2.1 [43] visualizes a moment diagram for selected families of distributions, where the chosen statistics to compare are the skewness γ_3 and the coefficient of variation, that is $\gamma_2 = \sigma/\mu_1$.

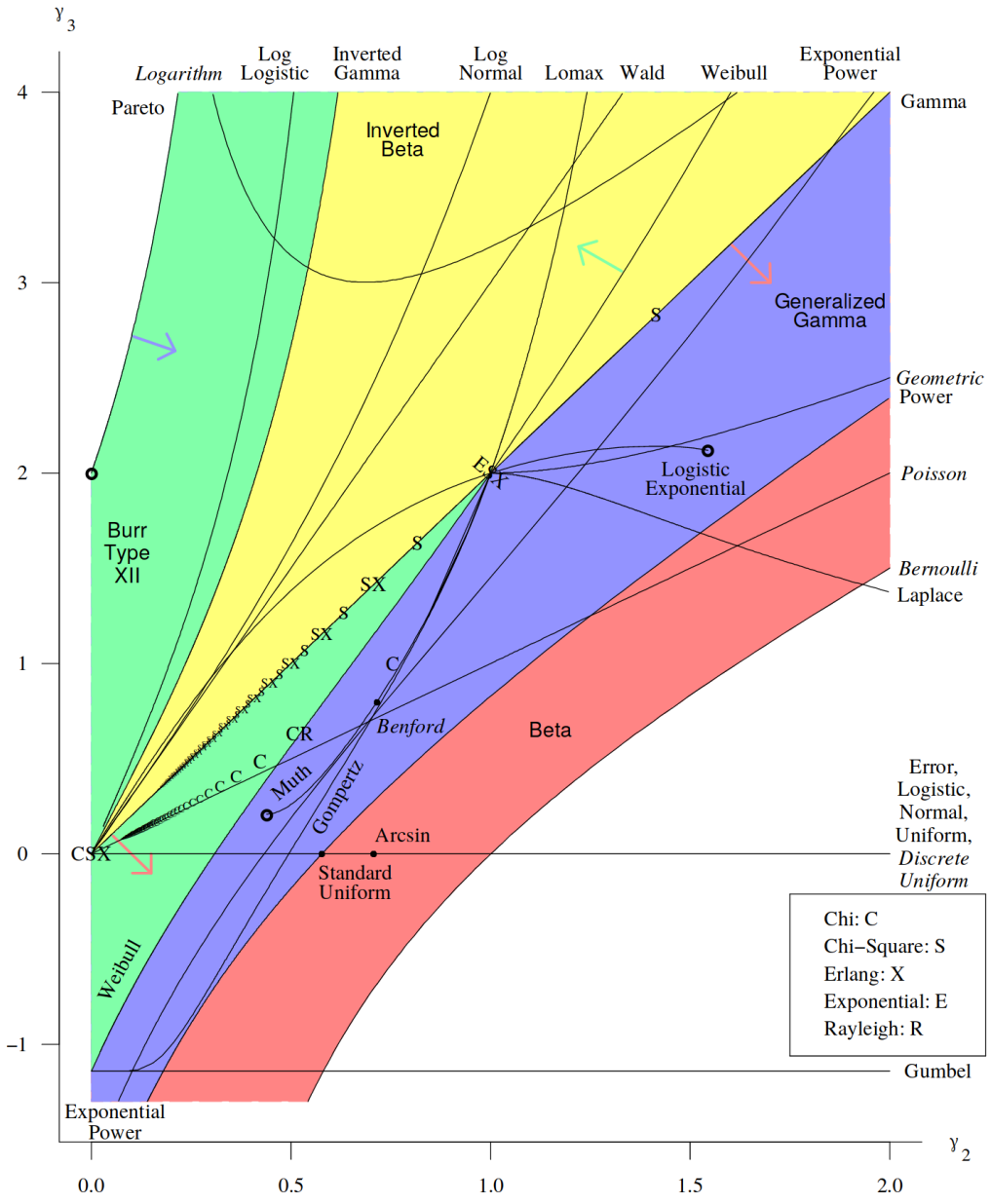


Figure 2.1: Moment diagram with various families of distributions for two statistic, the skewness and the coefficient of variation. The figure is provided by Vargo et al. in [43].

2.1.3 What Makes a Subpopulation?

The subpopulation structure for the variable of interest stands for what kind of patterns distributions of subpopulations have for the variable of interest. Subpopulations can be determined by a categorical variable, such as gender or age group, but in general a subpopulation can be any collection of data samples and not necessary determined by any variable. But when we wish to find patterns in the subpopulation structure, random segmentation is not a very fruitful approach. The problem of defining the subpopulation structure from the data requires finding appropriate subpopulations, which is not necessary a simple task, as the variable that would identify the subpopulations well can just be missing or using many variables can lead a better subpopulation segmentation than any single variable. Note that the concept of subpopulations is strongly present with mixture models in a similar sense we are using this term [29], [30], but the in-depth consideration of the structure of subpopulation distributions and its implications are often disregarded.

We illustrate the subpopulation structure concept in practice with an example in Figure 2.2 in which we display subpopulation distributions for 3 categorical variables. The variable of interest, for which the distributions are constructed, is the total weight. The 3 categorical variables in each of these cases are artificially created categorical variables and each variable has 4 discrete categories in this example. The variables are created by hand so that the first one is a random, second is dependent and third is determined by the weight variable. In the next example analysis we assume that we do not know the creation process behind the variables and just do the conclusions based on the looks of the distributions.

Example analysis of the subpopulation structure and its implications:

> In the case of the first variable the categorical variable seems essentially a random variable with no connection to the weight variable. All the distributions are almost the same. The third variable seems strongly connected to the weight variable. The distributions look very different for each subpopulation, and the variable total weight can be predicted based on the subpopulation category very strongly. The second variable seems dependent with the weight variable. The distribution shapes are slightly similar and the averages differ. Next we explain what these kind of structures imply for practical understanding and predictions about the variables used for making the subpopulations and about the variable of

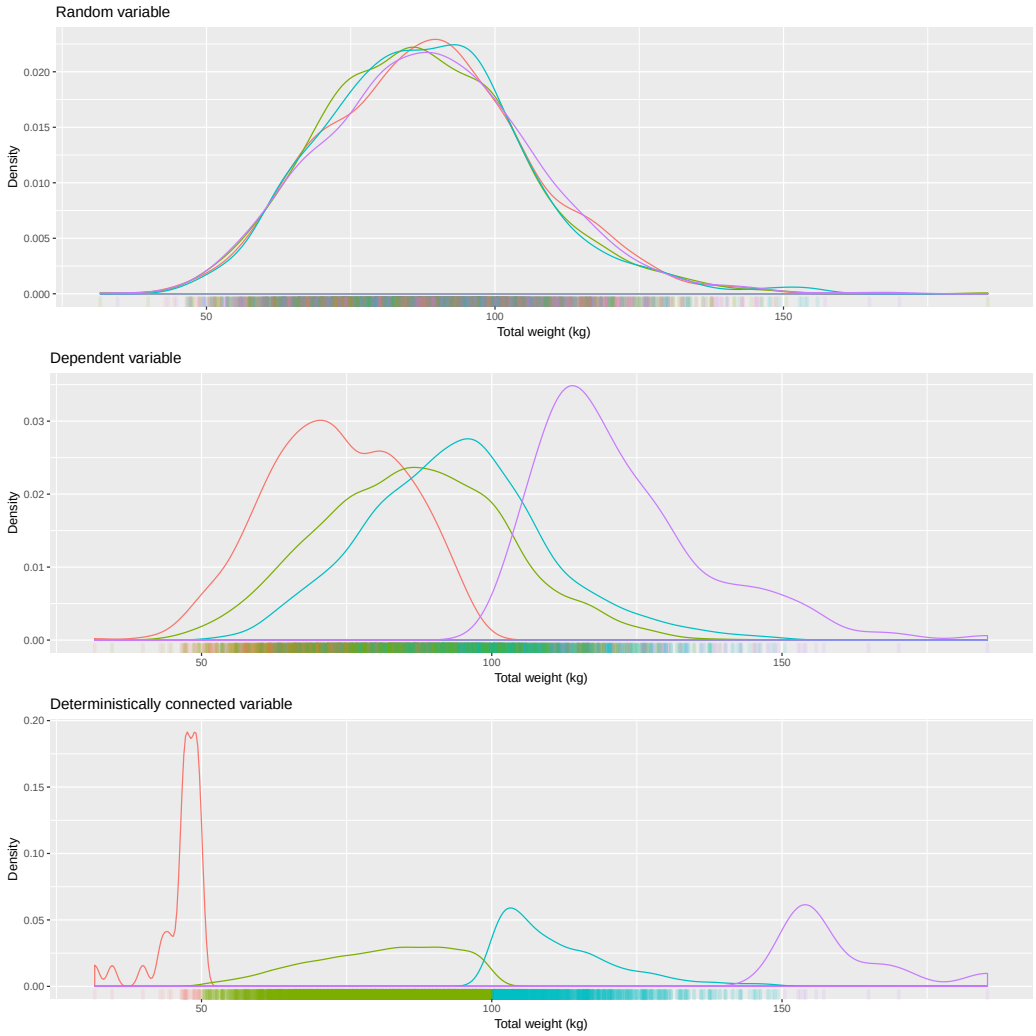


Figure 2.2: Subpopulations structures for different kind of subpopulations based on 3 categorical variables.

interest that is the total weight.

In the first case the variable used for categorization holds no information value by itself. However, if we would be introduced a fifth category, we could assume that it behaves like the known four categories and assume the common shape, that is the distribution shape for all samples, for that one too.

In the third case the variable is very useful for predicting the weight

variable, but there is no common distribution shape, other than it seem that the categories that have values under 100 kg are left-tailed and above it right-tailed. In case of estimating the distribution for a fifth category, we would have to rely on the shape of the category that has closest values to it, and ignore the information from the other categories, or just fit a normal distribution due to the lack of knowledge.

In the second case, we have some differences in the means and at least a little bit similarity in the shapes, that is we have both pros and cons from the first and third cases.

For our practical applications a reasonable wish is that:

- Each category should have as much differences in the means as possible, or some other criterion, such as differences is the deviations
- The distribution shapes should be as similar as possible

The first one enables that each segmentation is usable commercially and the second that we can expect the correct tail-behavior, extreme quantile estimates and such issues for even low sample groups, on top of having overall understanding. However, if we do not wish to have differences in the means in our particular application, but just to predict the shapes, it could be useful insight alone that the distributions are the same for all categories of certain variable, and alike we might only be interested in the differences in the averages, deviations or some other statistics for the subpopulations and be indifferent about the shapes.

Luckily, recognizable subpopulation structures seem to commonly exist in real life variables and subpopulations. See, for instance, many of the figures in Chapter 4 and 5 or any research, such as [24], were parametric families are fit to the data that potentially consist of subpopulations.

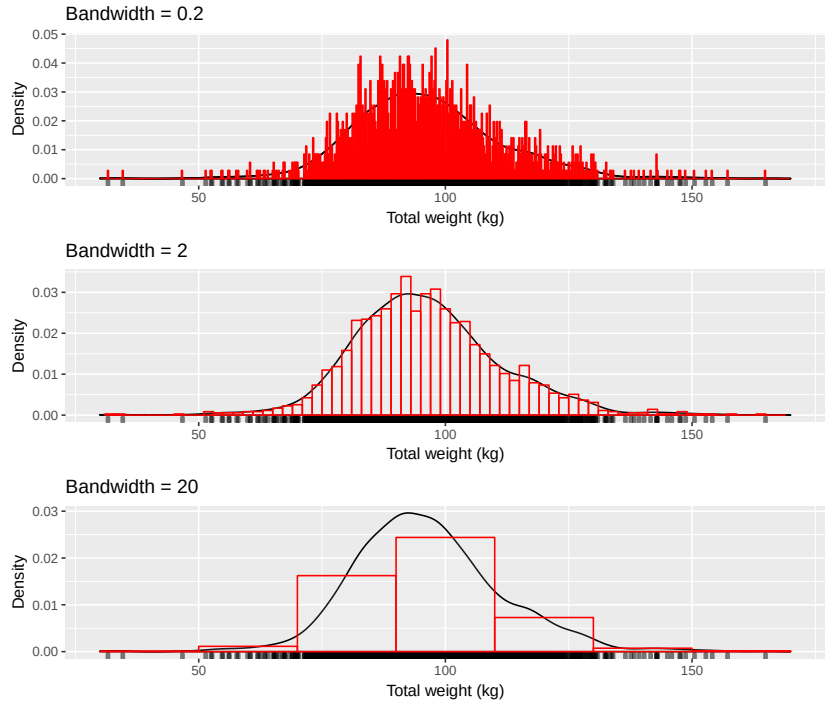
2.2 Density Estimation

In this section we cover some of the methods to estimate a probability density function for a subpopulation from a limited number of samples. The simplest method is to fit a histogram to the data. The kernel density estimation essentially is more advanced continuous version of the histogram fitting. We also examine practicalities of parametric fitting and finally mixture models.

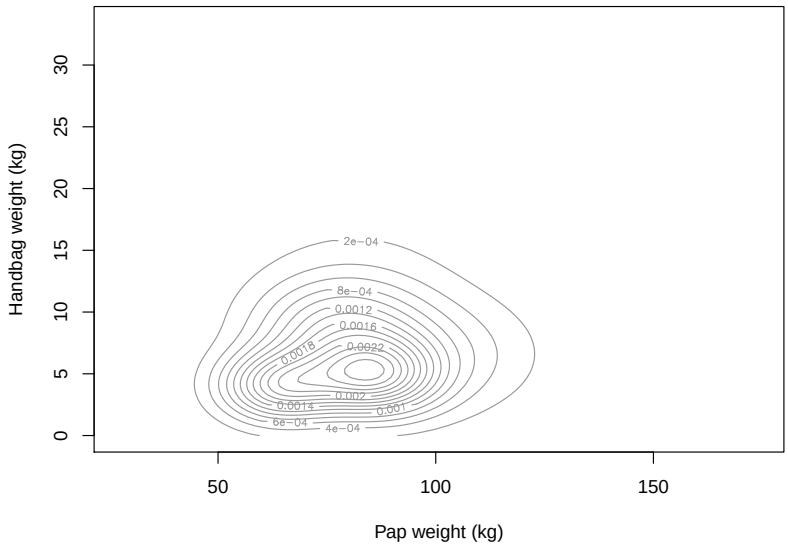
2.2.1 Histograms and Kernel Density Estimation

Estimating a probability density function from the data of continuous values, when no domain knowledge how the samples should be distributed is available, is not a trivial task. However, estimating the cumulative distribution can be done mechanically from the samples by the empirical cumulative distribution function, but which does not unfortunately provide estimates satisfactorily for many of the statistics of interest. Maybe the simplest way to estimate the probability density function is to fit a histogram to the data. This requires that the number of bins and their widths need to be chosen. There are a number of methods for choosing the bind sizes and the widths. In the upper part of Figure 2.3 we how histogram fits to the weight data. The middle one seems to have most applicable bandwidth in this case. Histograms can naturally be expanded to multidimensional settings. The lower part of Figure 2.3 shows a contour plot that is made for a two-dimensional histogram. Histogram methods are reviewed in more depth in the book of Silverman [40], which also examines the kernel methods that are discussed next.

More developed version of the histogram fitting is to fit so-called kernels to the data instead of discrete bar blocks. In this method called the kernel density estimation for each sample point a small distribution that is shaped accordingly to the selected kernel is fitted and then these distributions are summed together forming the final kernel estimate. Figure 2.4(a) illustrates how the final estimate is the sum of four little Gaussian distributions fitted to the samples. Figure 2.4(b) shows the effect of bandwidth that must be chosen for the Gaussian kernel. Choosing the correct bandwidth is a similar problem than with the histograms.

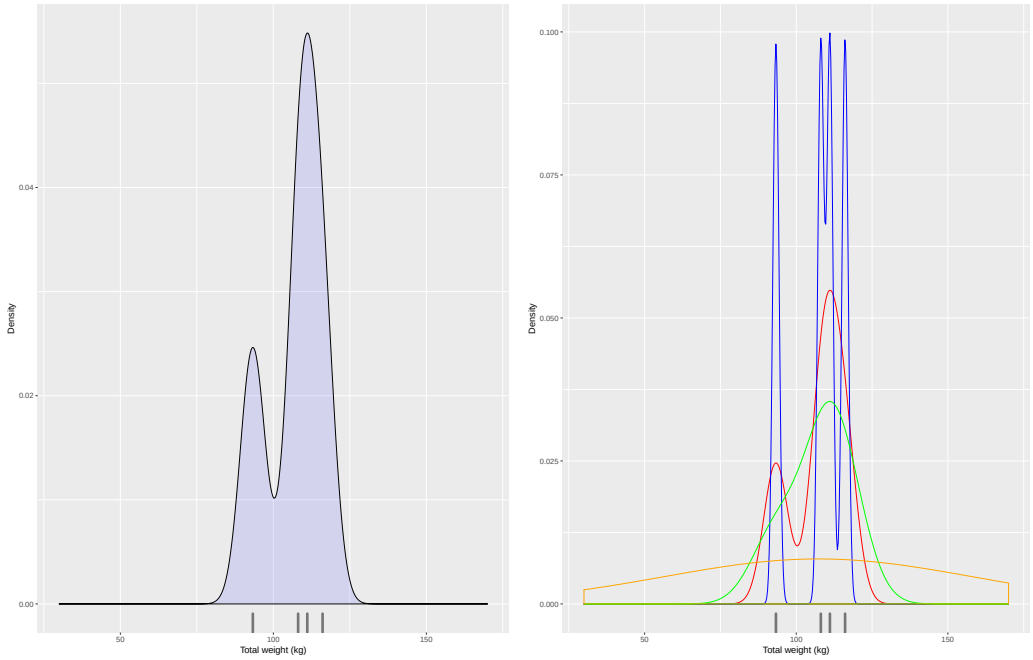


(a) Barplot histogram fits with different bandwidths to the weight data.



(b) The two-dimensional contour plot that shows the distribution of the pap and handbag weights.

Figure 2.3: Fitting histograms and contour plotting one and two-dimensional data.



(a) The default kernel density estimation fit to 4 samples using the automatically chosen bandwidth.

(b) Different kernel density fits with various bandwidths.

Figure 2.4: Kernel density estimation fits with different bandwidths.

Besides the Gaussian fit alternative kernel fits are shown in Figure 2.5. The type of kernel is particularly relevant with low sample size or when a certain tail behavior is desired. For instance, Gaussian kernel may provide more realistic prediction for large values as it does not reduce to nothing as fast as some of the other kernels.

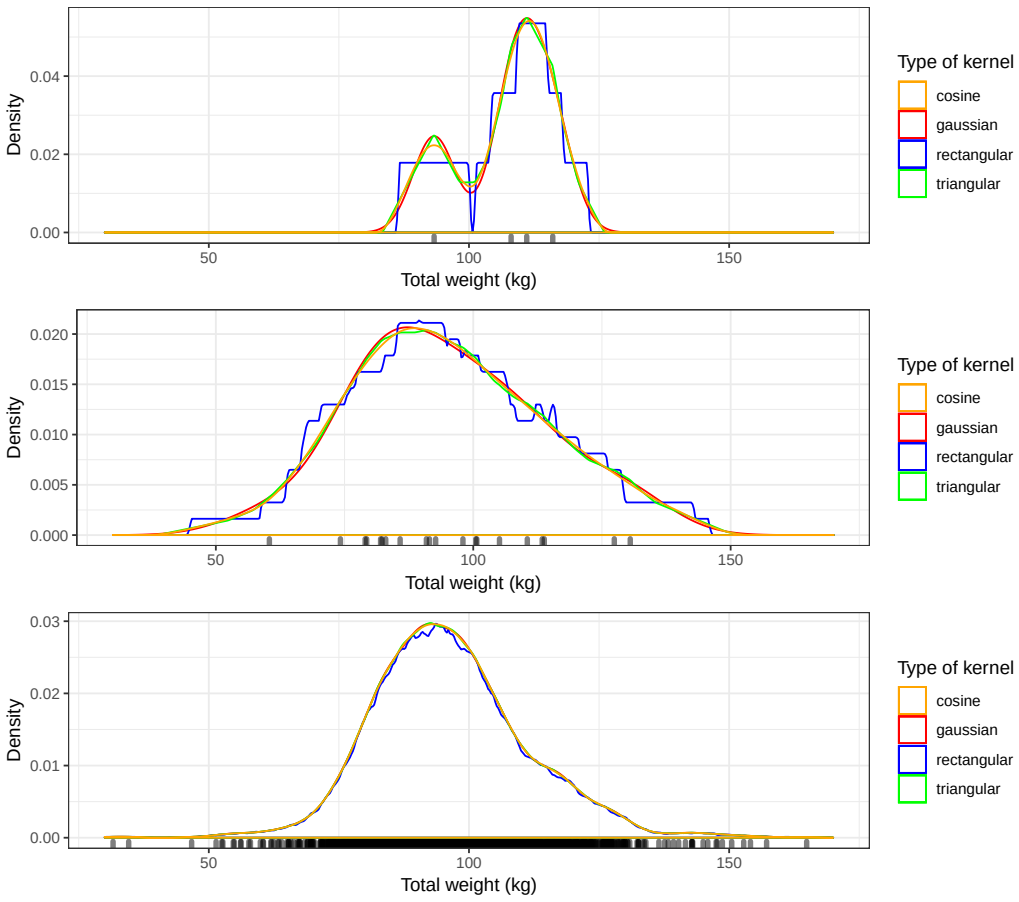


Figure 2.5: Kernel density estimation fits with different kernel functions with different sample sizes.

2.2.2 Parametric Methods

Parametric distribution fitting methods choose the optimal distribution out of a family of distributions that is known or believed to explain how the data is distributed. The optimal distribution is found by examining which parameters the data suggests according to some criterion. The parametric methods include, for example, maximum likelihood method and method of moments. When choosing the family of distributions for the parametric model, it should be considered whether symmetricity, unimodality, zero-valueness outside of a support or some other specific property is desired. Although the different parametric methods tend to give often very similar results, if a specific tail

behavior is wanted, the choice of method can make a great difference. In our calculations we use package *fitdistroplus* which documentation [8] also explains the aforementioned methods more thoroughly.

We explore how parametric fitting works to two different passenger weight subpopulation in Figure 2.6. The other has relatively many samples $N = 1773$ and the other only $N = 10$. Three fits, gamma, lognormal and normal Gaussian distributions, on top of the smooth fit, are done using the maximum likelihood method. Smooth fit refers here and in other figures to an appropriate kernel fit to the data.

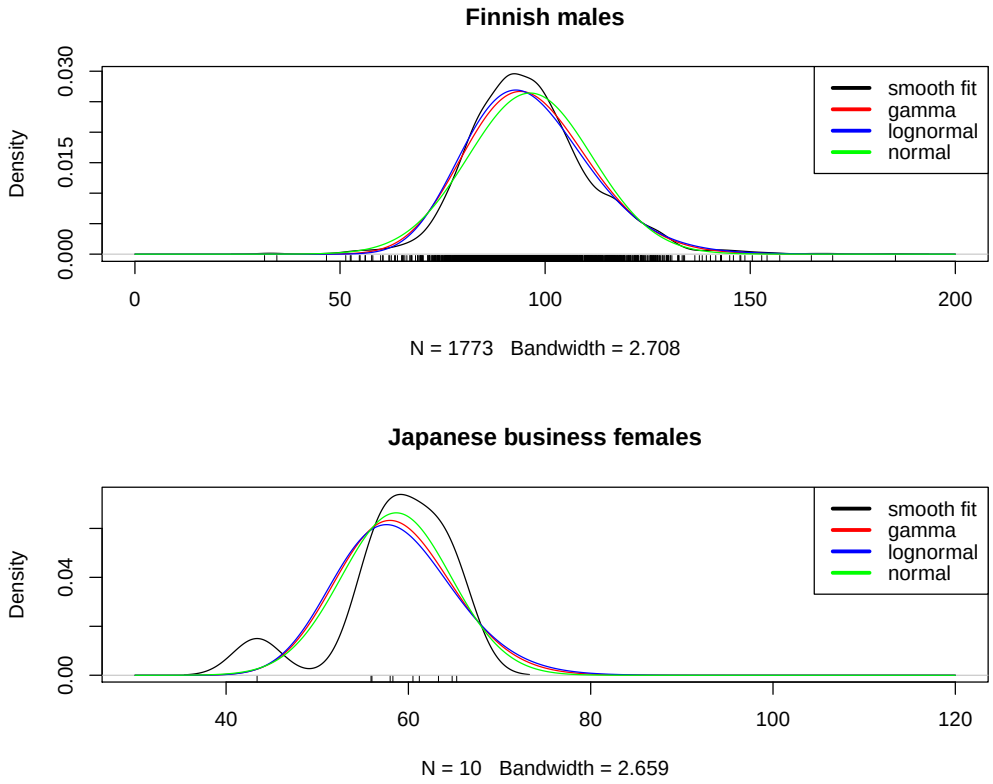


Figure 2.6: Parametric fits to two subpopulations in the weight data.

In the case of the Finnish males all the fits seem to align well with the true distribution. However, the perfectly symmetric normal fit fails to catch that the upper tail is longer than the lower tail and overemphasizes the very low weights. For the Japanese business females we can see that the smooth fit is probably pretty inaccurate due to low sample size and the parametric fits

reflect what the distribution would be if we had more samples, other than error in the tail behavior as the this sample is left-tailed, unlike the true distribution likely is. A validation shows that for these kind of passenger weight subpopulations the lognormal fit is the best, the gamma fit is the second, and the normal is the third in terms of the smallness of the error. Note that exponential fit and many other fits would be much worse, and thus, are not examined here.

There are also distributions that have have a large amount of parameters, which means that they can take the form of many different traditional distribution families. Examples of these include, commonly known Pearson type IV distributions and Johnsons's S_U-distributions, which use 4 parameter in their density functions. These multi-parameter distributions are not omnipotent though, and are usually unimodal, unless they are essentially a mixture of multiple distributions.

2.2.3 Mixture Models

Mixture models aim to explain the data by modelling the parts of it separately and then combining the model fits. [29] In practice, a collection of probabilistic models, such as Gaussian distributions models, are fitted to the data optimally in a mixture model. The mixture models and model-based clustering go hand in hand, the difference usually being the semantic end goal, whether the end objective is to find a distribution for the entire population or do the clustering to subpopulations with the fits. For a reference, see the next subsection 2.3.1 and an up-to-date article of mixture model developments [30].

We show possibly successful results of a mixture modeling process in Figure 2.7. The model seems rather robust as it gives similar results with all $N = 5090$ samples and only with around half of the samples $N = 2500$, whether the number of distributions is 2 or 4. In the best case scenario the mixture model recognizes subpopulations from the data, essentially doing what a clustering algorithm does. This kind of automatic clustering has, however, a few drawbacks.

Failing to fit the Gaussian mixture model in a satisfactory manner is demonstrated in Figure 2.8. In the case of fitting only two distributions, the model essentially fits one Gaussian distribution to almost all of the data and a very

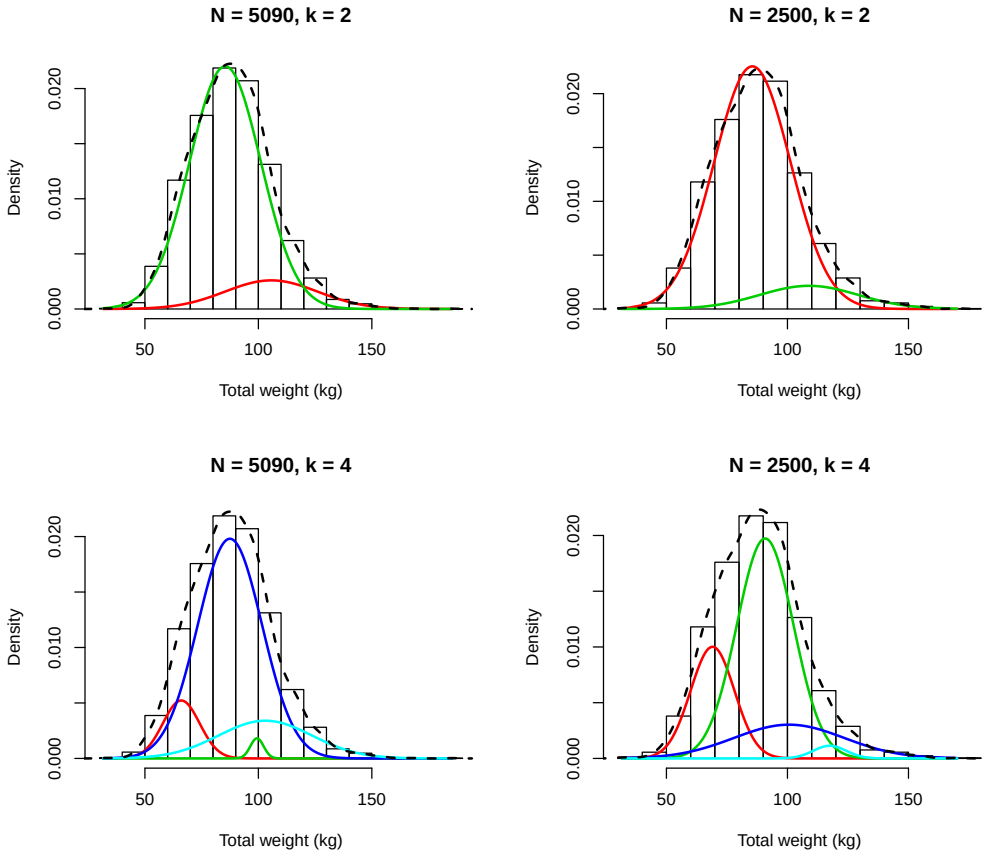


Figure 2.7: Examples of possibly good mixture model fits with different number of samples and selected number of subpopulations.

little distribution to very high weights, which gives too much density to extremely low values and having a subpopulation with mean that is over 140 kg seems unnatural. And in the case of fitting 8 distributions the model becomes unstable as removing just one sample completely changes the distributions. So the issues in general include, first of all, that it requires the choice of underlying models that are mixed. Secondly, that it can produce unrobust results when the number of desired groups is large. Thirdly, that the computation time to find the optimal fits, using the expectation maximization method as an example, grows inconveniently fast as the sample size and number of groups grows.

Due to these draw-back we can not utilize the mixture model directly to solve the all the problems for which the framework attempt to provide in-

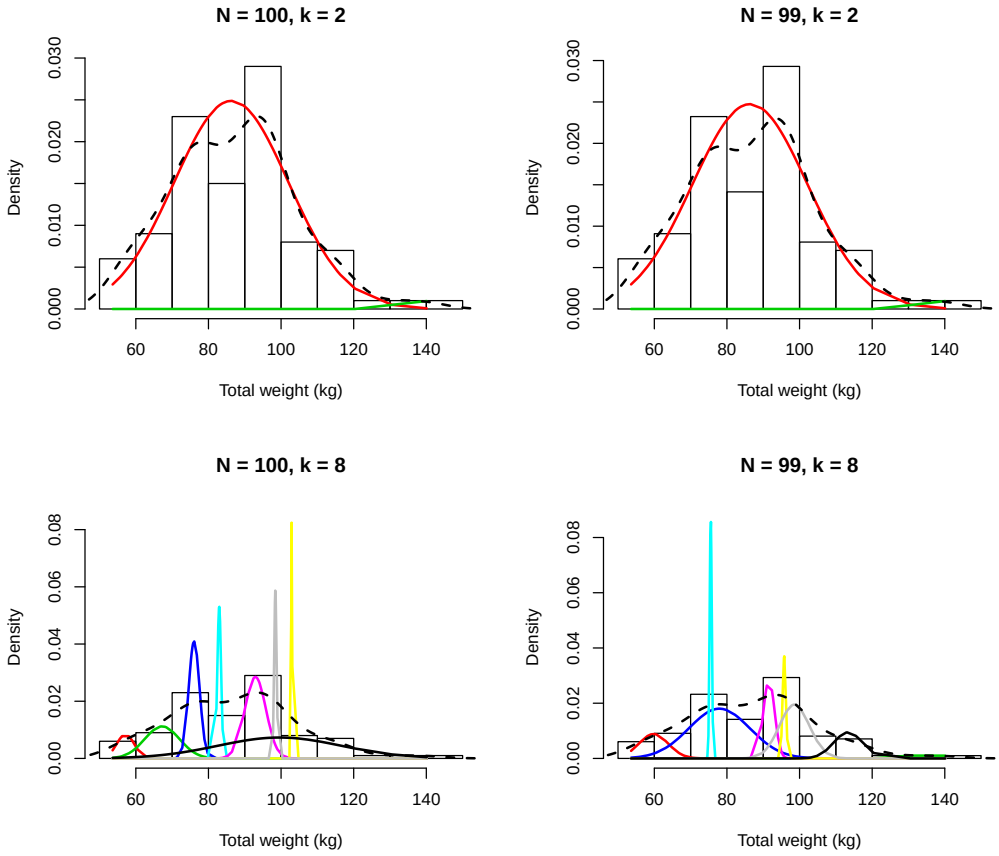


Figure 2.8: Examples that demonstrate the issues with mixture model fitting. Compare these fits to the previous figure 2.7.

sights. Nevertheless, the mixture modeling can be used when doing solely distribution fitting, such as in a case when it can clearly be seen that each subpopulation consist of two underlying distributions, which can not be separated using the actual clustering methods presented in the next section.

2.3 Subpopulation Clustering Methods

Clustering analysis provides a way to divide the data to more understandable segments and can possibly find subpopulations. Firstly, we consider model-based methods that share similarities with mixture models. But move on to consider methods that use the multivariate data instead of just the variable of interest, as secondly, the categorical group clustering is defined, that essentially uses other variables directly or with feature engineering. And thirdly, hierarchical clustering, a standard clustering algorithm for multivariate data, which has the potential to find the subpopulations, is explored.

2.3.1 Model-based Approaches to Clustering

In model-based clustering it is assumed that the data is generated by underlying model similarly than in the mixture modeling 2.2.3, see also again [30]. The usual approach is to find the parameters of the model by expectation maximization and assigning a data point to the cluster for which the likelihood of that point belonging there is the highest. [17] The model-based clustering aligns with our goal of finding subpopulations that are shaped accordingly to the same common shape structure, but it has similar drawbacks than the mixture modeling. The first one being that the underlying model needs to be chosen beforehand for subpopulations or at least some sort of assumption of the structure of subpopulations need to be made. [2] [11] [24] There are also attempts to do the clustering nonparametrically [1].

In Figure 2.9 we show example clustering to 5 groups using Gaussian distributions as the underlying model. See [12] for the implementation. The mean values of each differ desirably, but the second drawback can be seen at the borders of each cluster: the borders are too strict so that the subpopulation distributions would have common predictable shapes. Therefore, using model-based clustering as it is, would require that the points not in the closeness of the centers should be somehow fuzzily be shuffled to occasionally to be in a cluster other than the one with the highest likelihood. Also the problem that the model needs to be chosen beforehand could be solved by some iterative approach in which the framework is run multiple times and the distribution family, that is the underlying model, would be altered in the process. For the examples of this thesis, however, we do not formulate this

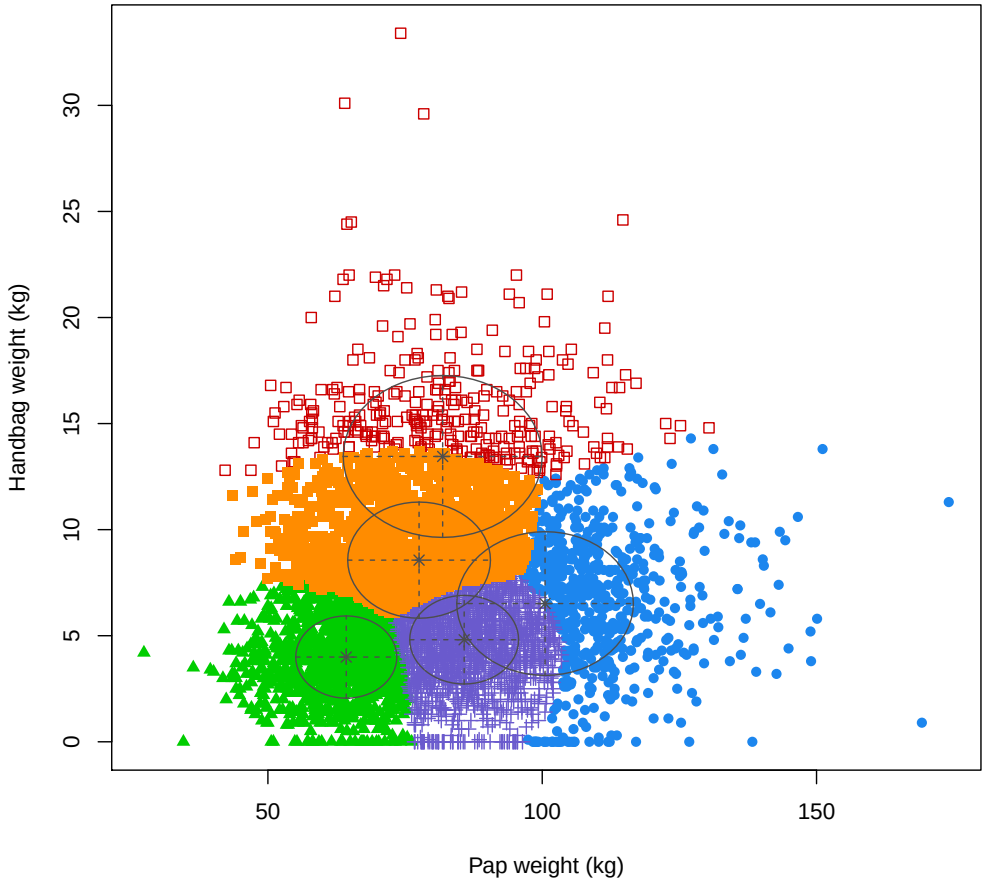


Figure 2.9: Model-based clustering to five subpopulations using the Gaussian distribution family fitted with the expectation maximization algorithm.

matter further, but give instead simpler and more tangible alternatives for the clustering that mainly use the multivariate data around the variable of interest.

2.3.2 Categorical Group Clustering

By the categorical group clustering, in its simplest form, we mean here that we use the most significant categorical variable in the data to group the data

so that each category forms a subpopulation of its own. Then we can further divide these categories using the second most significant categorical variable and so forth if wanted. A concrete example of this is given in Table 2.2. The term pivot table is also used for this representation in the context of many business intelligence and spreadsheet software applications.

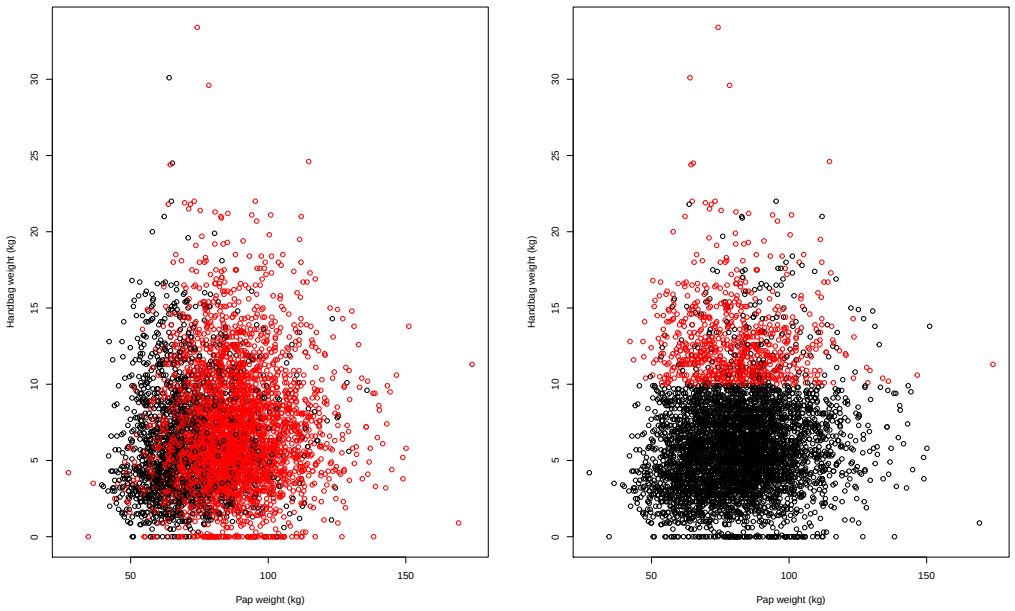
Table 2.2: Categorical group clustering using the variables gender, class and region with a few statistical aggregation values related to them.

	Gender	Class	Region	average	n	std	accuracy	confidence
1	Female.	Business.	atlantic	83.23	3	8.522	11.5861	9.6435
2	Female.	Business.	domestic	82.05	11	10.157	7.3162	6.0026
3	Female.	Business.	european	77.71	58	12.976	4.2973	3.3395
4	Female.	Business.	fareast	70.42	31	15.394	7.6957	5.4192
5	Female.	Economy.	atlantic	76.57	60	13.770	4.5505	3.4844
6	Female.	Economy.	domestic	77.42	235	13.682	2.2594	1.7493
7	Female.	Economy.	european	75.01	907	14.988	1.3004	0.9755
8	Female.	Economy.	fareast	69.49	235	13.619	2.5057	1.7413
9	Male.	Business.	atlantic	97.55	15	13.008	6.7484	6.5829
10	Male.	Business.	domestic	97.88	95	12.865	2.6431	2.5871
11	Male.	Business.	european	98.53	297	16.532	1.9082	1.8802
12	Male.	Business.	fareast	95.42	91	16.830	3.6241	3.4580
13	Male.	Economy.	atlantic	96.47	106	15.771	3.1121	3.0023
14	Male.	Economy.	domestic	95.66	666	14.439	1.1463	1.0966
15	Male.	Economy.	european	94.07	1756	15.558	0.7736	0.7277
16	Male.	Economy.	fareast	86.83	524	16.092	1.5868	1.3778

The determination which variables are the most significant can be done in many ways, but the variable selection problem is challenging in general. In machine learning the process is called feature selection. One practical alternative is to use linear regression model to explain the variable of interest. Random forest methods, using domain knowledge or just testing out different combinations and picking the best clustering using cross-validation in the end, may also work. See, for example, *caret* package [23], which provides a systemical procedure for variable selection. Furthermore, we can try create more features based on the variables available and possible domain knowledge. [16] Also the particular subpopulation of interest should be considered, and picking the variables just based on the the subpopulation of interest is often the best way to go if accurate estimate values are wanted instead of overall understanding of the data.

Categorical group clustering has a number of simple reasonable feature engineering modifications that make the method more powerful, but also make picking the optimal clustering more challenging, for instance:

- The numerical variables can be transformed into categorical ones
- The alternative of not using every category for a categorical variable, but combining the categories in it
- Only dividing a certain category in a variable further instead of all categories in it



(a) Using the variable gender that has two values.

(b) Using the variable that implies whether passanger is in class economy and that the handbag weight is heavy.

Figure 2.10: Categorical group clustering results using different variables in the weight data.

Categorical group clustering works well often just by choosing the most significant variables and combinations of them, if the variables in the data are of good quality. Nonetheless, not all categorical variables necessary have that much information value or the distribution formed can have little common

structure, which means that our understanding about the nature of the data will not be very usable. Also when doing transformations from the numerical variables to categorical ones extra care should be taken so that distribution shapes are understandable. An example clustering using the gender variable is shown in Figure 2.10(a). Notice how the borders of each clusters are much more fuzzy and it is more believable compared to model-based clustering 2.9 that these groups would be from some similarly shaped distributions. We also show in Figure 2.10(b) an artificially created variable that indicates whether passenger is class economy and has a heavy handbag. These kind of variables that are directly connected to the variable of interest are problematic in a sense of wanting to preserve the common distribution shape. We take a more in-depth look into the distribution shapes in Section 4.3.1 using the categorical group clustering.

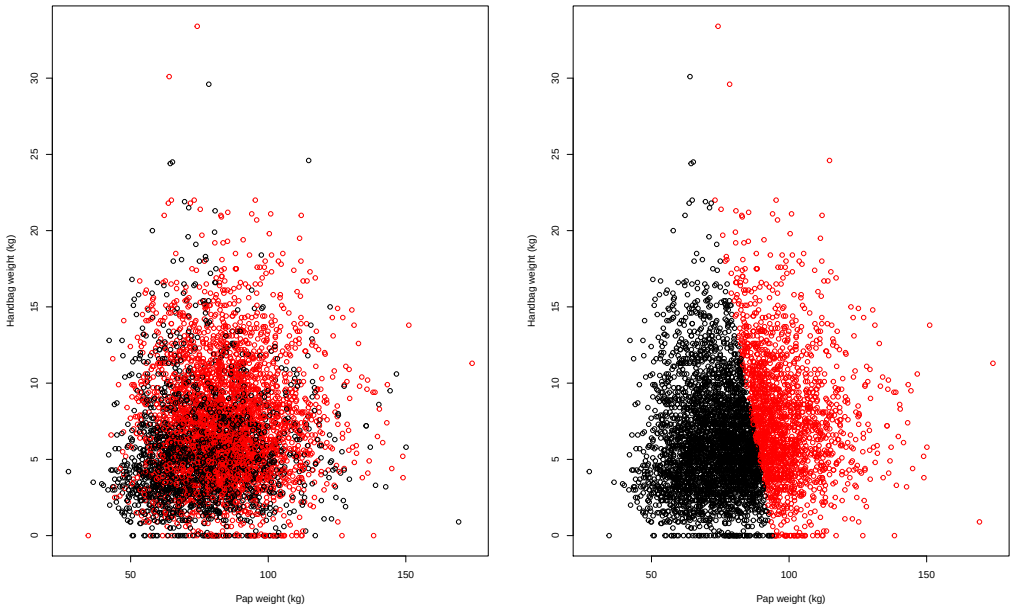
2.3.3 Hierarchical Clustering

Hierarchical clustering is a standard clustering method in which each data point is initially its own cluster and then at each stage of the clustering the two most similar clusters are joint together until in the end there is only single cluster. The aforementioned clustering method is called the agglomerative bottom-up approach, which we will refer from now on as just hierarchical clustering., but there also is the divisive top-down approach, that starts with one cluster which is divided iteratively unlike merging that happens in the agglomerative approach. The hierarchical clustering allows us choose the cut-off value, that is how many clusters we want, afterwards. [17]

The core of this algorithm is the cluster similarity. Mathematically, we must choose a linkage criterion that determines the distance of two clusters and a metric affecting the shape of the clusters that is used in the linkage criterion. For example, we can use the maximum linkage $\max\{\, d(x,y) : x \in X,\, y \in Y \,\}$, where X and Y are clusters, and $d(\cdot)$ is the metric that can be, for instance, the Euclidean distance $d(x,y) = \|x - y\|_2 = \sqrt{\sum_i (x_i - y_i)^2}$, if we have a continuous numerical variables. The metric can be also defined for categorical and other type of variables.

In our case, the metric for the mixed multivariate data can be, for example, the Gower distance [14], that is sort of a jack of all trades metric, that can take numerical, categorical, logical and ordinal variables simultaneously and also allows the weighting of variables according to their importance. In short,

the Gower distance computes the weighted average of dissimilarities for each variable, where the dissimilarity means for numeric variables the absolute difference divided by the range of that variable, for categorical the distance is 0 or 1 depending whether their value is the same and in similar fashion for the other type of variables. The weights in the weighted average can be chosen to be anything non-negative.



(a) Using all the categorical variables in the weight data.

(b) Using only the total weight variable.

Figure 2.11: Hierarchical clustering results using different variables in the weight data.

Examples of the hierarchical clustering are shown in 2.11. The used metric is the Gower distance. The first figure 2.11(a) shows the hierarchical clustering using every categorical variable in the weight survey data. The results resemble the gender based categorical clustering results. The second figure 2.11(b) shows the clustering when it is done based on the total weight variable. This clustering resembles solution to a traditional clustering problem without the consideration of subpopulation distributions and shows that using variables directly connected to the variable of interest is also problematic with the hierarchical clustering. So, one of the differences between the categorical clustering is that the latter requires more domain knowledge to use, but using the hierarchical clustering blindly to every variable is not a good practice

either. We return to further examples of using the Gower distance with the average linkage and various weightings in the examples of Chapter 4, but before this we finalize our framework by discussing how common shape can be identified after using the subpopulation clustering and distribution fitting methods that were showcased in this chapter.

Chapter 3

Common Shape Identification and Subpopulation Estimation

In this chapter we develop simple algorithms for shape identification under a few statistical assumptions as well as discuss ideas for more complex methods and how to choose subpopulation similarity weights in the algorithms. In other words, we define how the third and forth block of the framework could be implemented, that is how we obtain the improved estimate for the distribution of the subpopulation of interest.

3.1 Location-Scale Averaging Method

Here we build a simple way to identify the common shape of subpopulations assuming they are from the same location-scale family. However, the method may work decently well even if the location-scale assumption does not hold. Similar approaches have been used in other studies to determine shape similarity between distributions, although the idea of translating and scaling back to the original values is not present in these methods. Wegner et al. [45] use similar normalization method to ours to determine whether two graphs are similar and Osada et al. [33] to measure the similarity of shapes of three dimensional physical objects.

In location-scale averaging method the common shape is identified by calcu-

lating the translated and scaled distributions, that is normalized versions, out of all subpopulations and then calculating the weighted average out of those. After the common shape is found, that is the weighted average, the distributions are translated and scaled back to the original locations and scales. The process is mathematically formalized in Algorithm 1 and visualized in Figure 3.1.

Algorithm 1 Location-scale averaging method:

1: $C_1, C_2, ... \leftarrow ClusteringAlgorithm(X)$
▷ Cluster the multivariate data using some clustering algorithm
2: $Z_1, Z_2, ... \leftarrow (C_1 - mean(C_1))/sd(C_1), ...$
▷ Normalize the data based on the variable of interest for every point in each cluster
3: $Z = Merge\ Z_1, Z_2, ...$
▷ Merge all scaled clusters together
4: $C_1^*, C_2^*, ... \leftarrow Z * sd(C_1) + mean(C_1), ...$
▷ Scale and translate the common shape back for every point in each cluster
5: $D_1^*, D_2^*, ... \leftarrow DistributionFittingAlgorithm(C_1^*), ...$
▷ Fit distribution using some distribution fitting algorithm

The left part of Figure 3.1 shows 7 subpopulations determined from the data, which seem to have a similar shape. However, it can be seen that some of them have irregularity mostly due to a low sample size. In the middle the standardized versions of the subpopulations can be seen. The average common shape is the weighted average. There are many ways to assign the weights for each subpopulation, such as the sample size of each subpopulation, which is used here. The right part shows the distributions after translating and scaling back to the original place and scale. For high sample groups the shape changes only a little, but the highlighted red subpopulation gets its shape corrected significantly.

The method has a sound intuition behind it, but a deeper look at the theoretical exactness, convergence and possible faults should be considered. The first matter is to confirm that the distributions provided by the algorithm have the same expected values for relevant distribution statistics as the trivial estimate with a finite number of samples.

Now, let $X_1, X_2, ..., X_N$ be random variables from the same location-scale family and $C_1, C_2, ..., C_N$ be collections of random samples drawn from these random variables and $C_1^*, C_2^*, ..., C_N^*$ be collections of points calculated by the location-scale averaging method, $D_1, D_2, ..., D_N$ distributions fitted to

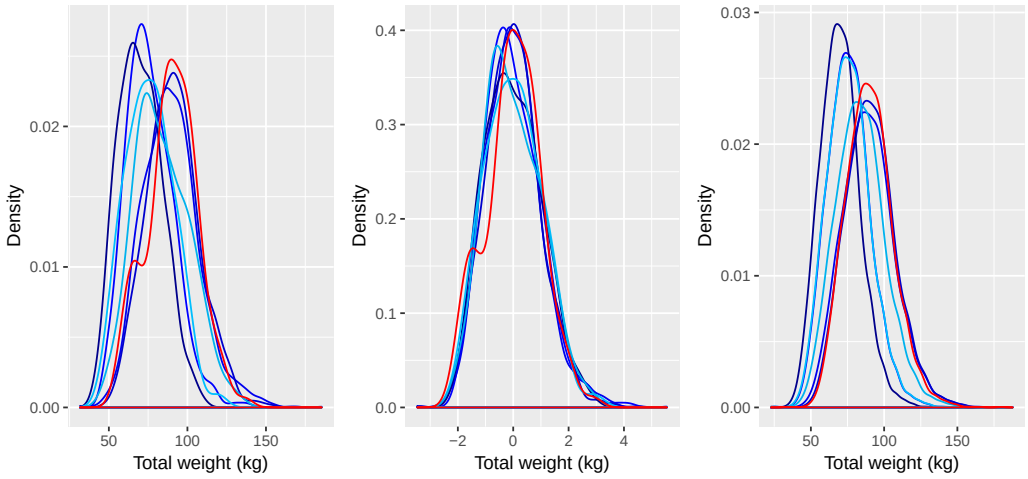


Figure 3.1: The location-scale averaging method step by step.

$C_1, C_2, ..., C_N$ and $D_1^*, D_2^*, ..., D_N^*$ distributions fitted to $C_1^*, C_2^*, ..., C_N^*$. It should hold that the expected values of relevant distributions statistics are the same. By relevant statistics we mean moments and other quantities that relate to the distribution location, scale and shape, but not for instance sample size related statistics like the number of samples. Denoting the function which calculates the distribution statistics with G, it should hold for any subpopulation of interest C_I that $\mathrm{E}\left[G(C_I)\right] = \mathrm{E}\left[G(C_I^*)\right]$.

For instance, for the mean value it should hold that

$$\mathrm{E}\left[mean(C_I)\right] = \mathrm{E}\left[mean(C_I^*)\right]. \tag{3.1}$$

To verify this, let N_i be the number of samples in C_i and $N_{total} = N_1 + N_2 + ... + N_N$ is the total amount of samples, then expanding the right side of this equation gives:

$$\begin{aligned} mean(C_I^*) &= mean(Z * sd(C_I) + mean(C_I)) \\ &= \sum_{p=1}^{N} \left[\sum_{k=1}^{N_p} \left[\frac{C_{k,C_p} - mean(C_p)}{sd(C_p)} * sd(C_I) + mean(C_I) \right] \right] / N_{total} \end{aligned} \tag{3.2}$$

Now we see that

$$\begin{aligned}
&\sum_{p=1}^{N}\left[\sum_{k=1}^{N_p}\left[mean(C_I)\right]\right]/N_{total} = mean(C_I) * \sum_{p=1}^{N}\left[\sum_{k=1}^{N_p}\left[1\right]\right]/N_{total} \\
&= mean(C_I) * \sum_{p=1}^{N}\left[N_p\right]/N_{total} = mean(C_I) * N_{total}/N_{total} = mean(C_I)
\end{aligned} \tag{3.3}$$

and also that

$$\begin{aligned}
&\sum_{k=1}^{N_p}\left[\frac{C_{k,C_p} - mean(C_p)}{sd(C_p)}\right] = \sum_{k=1}^{N_p}\left[C_{k,C_p}\right]/sd(C_p) - \sum_{k=1}^{N_p}\left[mean(C_p)\right]/sd(C_p) \\
&= \sum_{k=1}^{N_p}\left[C_{k,C_p}\right]/sd(C_p) - N_p * mean(C_p)/sd(C_p) \\
&= mean(C_p) * N_p/sd(C_p) - N_p * mean(C_p)/sd(C_p) = 0, \\
&\forall p \in 1, 2, ..., N.
\end{aligned} \tag{3.4}$$

Substituting 3.3 and 3.4 to 3.2, we get that $mean(C_I) = mean(C_I^*)$, which directly implies that 3.1 holds. Similar verification should be possible for the variance and the statistics related to the higher moments and we could also try to verify that $\mathrm{E}\left[G(C_I)\right] = \mathrm{E}\left[G(C_I^*)\right]$ is equal to the theoretical statistic $G(X_I)$ that random variable X_I has, but in this case the location-scale assumption, the expected value operator and sampling from the random variables should be used in the verification. It should be noted that if we also consider the distribution fitting parts of the method matters, like the chosen kernel function, affect the estimates. For example, a relatively large bandwidth may increase the deviation of the estimates.

We can make hypotheses how the algorithm is assumed and should work, although we will not prove these statements and it is even possible that these statement do not hold at least under some special cases. Here we assume that the random variables $X_1, X_2, ..., X_N$ are from the same unknown location-scale family. And when referring the convergence, we assume that $N_{total} \to \infty$. The center part, that is disregarding the clustering and distribution fitting, should have the following properties:

- $\mathrm{E}\left[G(C_I^*)\right] = \mathrm{E}\left[G(C_I^*)\right] = G(X_I)$

- Reasonable estimators calculated for $C_1, C_2, ..., C_N$ converge faster when calculated for $C_1^*, C_2^*, ..., C_N^*$ and also converge in probability to $G(X_I)$
- Weighting by the number of samples is the best weighting in the terms of convergence of reasonable estimators

And the entire location-scale averaging method should also have the following properties assuming a reasonable distribution fitting method is used:

- $D_1^*, D_2^*, ..., D_N^*$ convergence in distribution to the same as $D_1, D_2, ..., D_N$ converge, which is equal to probability density distributions $f_1, f_2, ..., f_N$ of random variables $X_1, X_2, ..., X_N$
- The earth mover's distance $EMD(f_I, D_I^*)$ converges to 0 in probability and the rate of convergence is at least as fast as $EMD(f_I, D_I)$

The location-scale averaging method can be improved especially, when location-scale assumption holds only weakly or a certain subpopulation just does not follow the common shape in the reality. In this case the smooth fit can be more accurate estimate than the location-shape method estimate. Particularly if the sample size for the subpopulation of interest is large, the smooth fit should be trusted with a high confidence. In general, we can combine the predictive accuracy of the smooth fit to the location-scale method using a weighting function that assess our trust in the smooth estimate versus the location-scale method. The weighting function gets a weight value w between 0 and 1 based on the number of samples N in the whole data set and the number of samples N_i in the subpopulation i that is considered. An example of a reasonable weighting function is

$$w_i = \frac{N_i}{(1-\alpha) * N_i + \alpha * N}, \tag{3.5}$$

where α is a further constant between 0 and 1 to model our trust whether we expect the subpopulation of interest to be similar to the rest of subpopulations. If we can construct a validation data set, we could also determine our trust in the two models by using emsemble methods [32] that are commonly used to combine multiple machine learning methods. The location-scale averaging method mixed with the smooth fit is described in Algorithm 2.

Algorithm 2 Location-scale averaging method mixed with the smooth fit:

1: $D_1, D_2, ... \leftarrow DistributionFittingAlgorithm(ClusteringAlgorithm(X))$
▷ Fit distributions to clustered data
2: $D_1^*, D_2^*, ... \leftarrow LocationScaleAveragingAlgorithm(X)$
▷ Fit distributions using the location-scaling
3: $w_1, w_2, ... \leftarrow SmoothWeighting(N_1, N, \alpha), ...$
▷ Calculate the weightings for models
4: $D_1^{*s}, D_2^{*s}, ... \leftarrow w_1 * D_1 + (1 - w_1) * D_1^*, ...$
▷ Mix the results of the methods

3.2 Approaches of Higher Moments

The approach in the previous section understood the shape by using the first two moments. One of the benefits of this is the access to easy standardization of distributions. When trying to understand the shape using the higher moments than the first two, we do not have a well-known family concept similar to the location-scale family and there are no easy tricks to do the standardization in general. One specific way to do the standardization with respect to the 3rd moment is to use the Box-Cox transformation which removes the skewness, and so the 3rd moment. In general, the power transformations are a technique to make the data distributed more like the normal distribution [4], for which we could apply location-scale method like standardization techniques. To use these kind of methods we would need to determine for which moments we do the standardization in the first place. As such, we do not try to do the standardization using the higher moments here, but try to find other kind of common moment patterns to access the underlying common shape.

Moment diagrams serve as our main tool to understand the patterns in the distributions. Once the prominent moment patterns have been recognized visually, the patterns need to be used when fitting the distribution for the target subpopulation. We take a look at a couple of examples. Multiple moment diagrams for various centralized moments are shown in Figure 3.2 for all flight-gender subpopulation combinations with respect to the total weight variable. The points are not distributed randomly as there are some notable patterns in these charts. However, the genders that are shown in different colors are hard to separate which is expected as they have similar distribution shapes in the data.

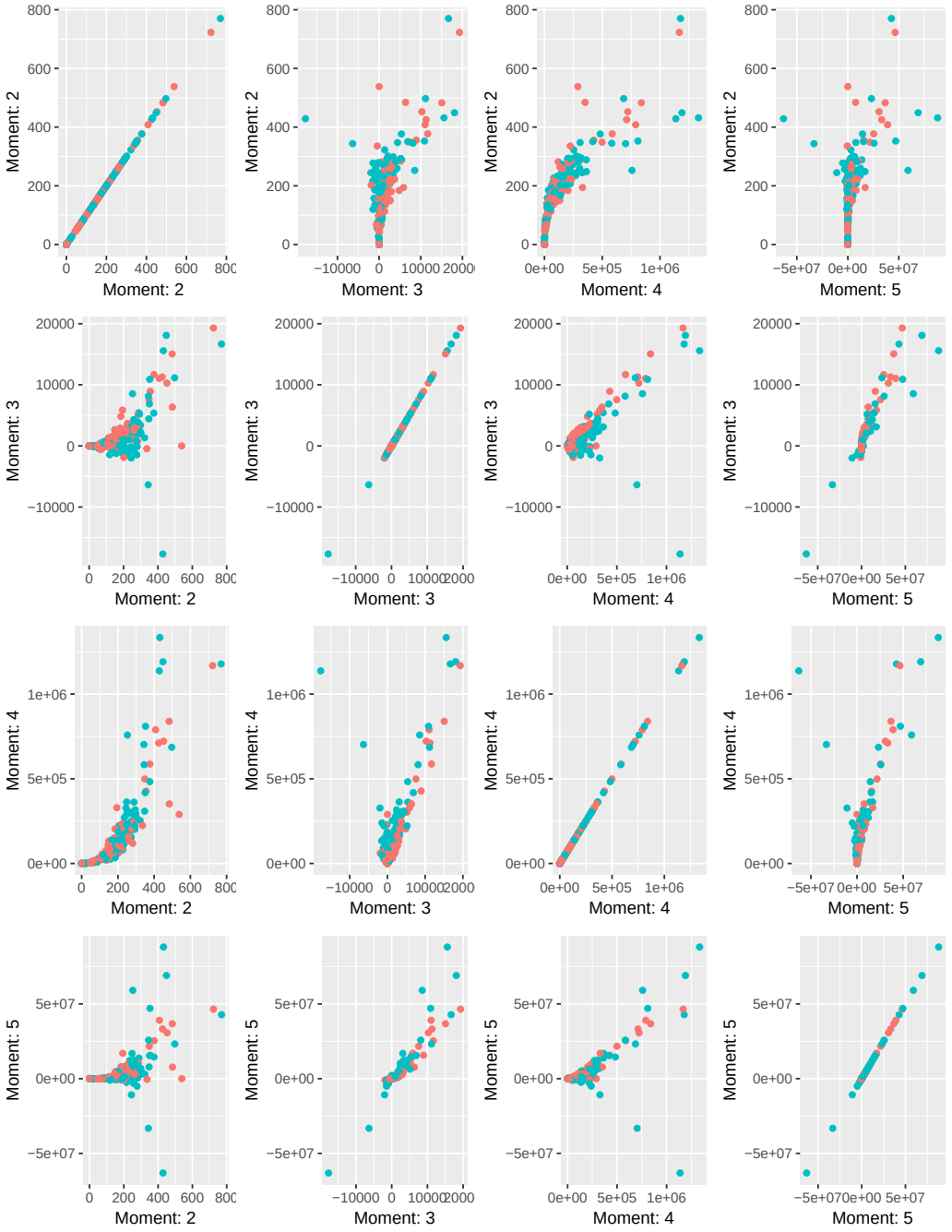


Figure 3.2: A diagram of the centralized moments for all flight-gender subpopulation combinations. The genders are shown in different colors. The variable of interest is the total weight for all points.

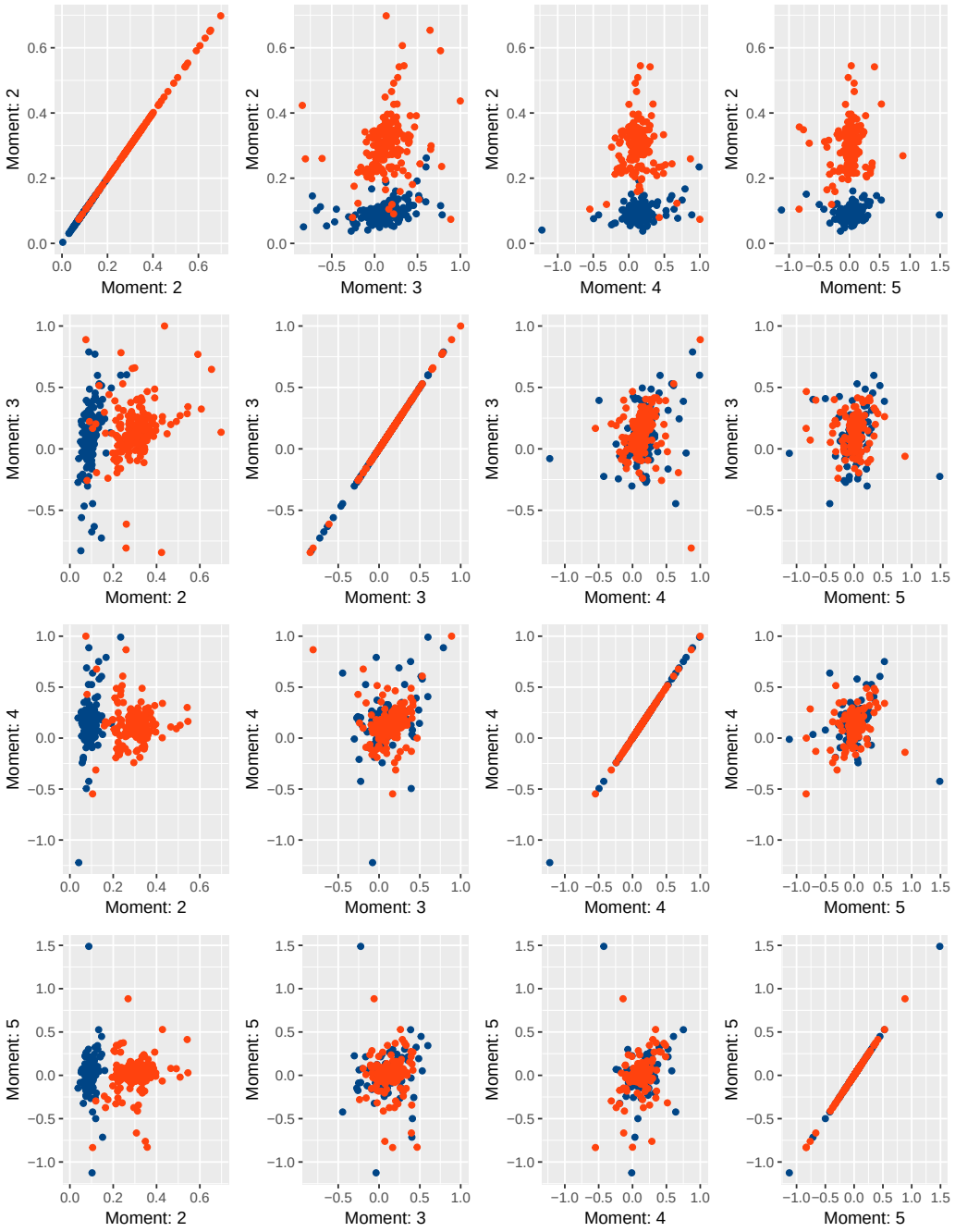


Figure 3.3: A diagram of the centralized L-moments for all flight-gender subpopulation combinations. Two variables of interest are considered, the total weight and the handbag weight, these are shown in different colors.

Figure 3.3 shows similar visualizations for centralized L-moments for two different variables of interest, the points still being all flight-gender subpopulations. The variables are the total weight and the handbag weight and they are shown in the different colors. This time the diagrams show less predictable patterns within a subpopulations, but the diagrams with the second moment axis make it possible to distinguish the total weight and handbag distributions, which is expected as the distributions shape is a bit different for the total and handbag subpopulation distributions, unlike male and female distributions that were considered in Figure 3.2. These examples support the hypotheses expressed in 2.1.2 about possible identifiable moment patterns for subpopulations and provide basis for more advanced methods to understand the shape similarity.

Next we describe the outlines for two different approaches to utilize the diagrams algorithmically.

Outline of the method of regressional moment patterns:

> The idea of this method is to fit some general purpose distribution of many parameters to the subpopulation of interest using the method of moments, but instead of calculating the moments directly from the subpopulation, they are evaluated using the moment diagrams patterns. That is, we determine which moments or distribution characteristics pairs have noticable patterns in them using the moment diagrams. Then we fit regression lines, or curves in more advanced settings, to the diagrams. Each moment used in the method of moments is calculated from the regression lines, by calculating the other moments directly from the other subpopulations and then the moment of interest is predicted from the regression line for the subpopulation of interest. If the same moments are calculated from multiple diagram pairs the value can be the weighted average, where the weights are determined by the coefficients of determination. Similarly to the location-scale averaging method, these higher moment methods can naturally be mixed with the smooth fit estimate similarly to Algorithm 2.

Outline of the matching moments by random sample addition method:

> In this method we try to modify the original subpopulation sample set by adding new points so that the moment requirements are fulfilled better. The method requires similar pattern exploration from the moment diagrams as the previous method. After this is done random points are

added to the subpopulation of interest, so that the moment requirements are met better. The adding of the points will lead to unwanted results if done naively, so we do the following. First, multiple randomly selected points should be added at once, instead of just one. Secondly, to find the optimal points to be added, a large enough number of trials should be made. Thirdly, these steps should be repeated so that enough points are generated for which the smooth fit can be finally made. On top of these, the moments to be matched must be based on strong patterns or the method will not work at all.

We need a substantial number of subpopulations with large amounts of samples and extensive manual research of moment diagrams to these methods to work in general case. In practice, we may, however, implement something more simple which uses the principles described above, but assumes that distribution is relatively uncomplicated, such as being unimodal. In this case, for instance for the passenger weight data subpopulations, the improved estimate can be calculated relatively successfully as follows, the results shown in Figure 3.4:

- Calculate the first two moments directly from the subpopulation of interest
- Calculate the third and fourth moment using linear regression in a few moment diagrams that have third and fourth moments in them
- Determine the parameters in Pearson type IV distribution, see 2.2.2, using the calculated four moments

In general, generating a distribution unparametrically based on its moment is a variation of the problem of moments, which has been studied, for instance by John et al. [22], who moreover consider the case where only a finite number of moments are known and the sample moments are calculated from a limited number of samples, which is our case also. But for now we refrain from further considerations of these more advanced methods, which is rather reasonable as the tools discussed here are more than enough for our case examples. We proceed to summarize the practicalities of using the framework next.

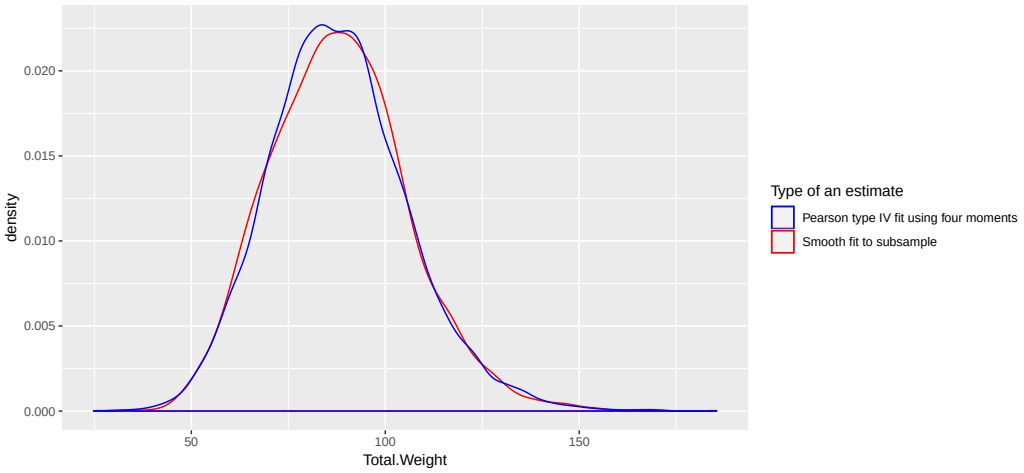


Figure 3.4: A Pearson type IV fit using four moments to the weight data.

3.3 Practical Notions on Estimates and the Framework

The methods examined in the previous sections provided algorithms that may work in theory, but in practice very simple models may outperform more complicated ones. In our case, it often is the case that instead of using all subpopulations, using just the subpopulations that are known to be similar to the subpopulation of interest should be utilized or at least weighted more. Which subpopulations to use could be based on the information provided by the moment diagrams as discussed. Other alternative is to use expert knowledge. The alternatives of this include that a expert recognizes the subpopulations directly or by identifying the most significant variables which are also believed to have subpopulation structure within them. Also if the clustering is done using the categorical clustering, we get a natural way to find potentially similar subpopulations. A simple example of this with the weight data is that subpopulations with similar demographic properties are weighted more. For example, if we were to understand the weights of Atlantic male passengers, we could cluster the population to region-gender subpopulations and assign a greater weight to all Atlantic and male subpopulations than to the other subpopulations.

The framework should be remembered to be modified according to the data and the operational interests. Discrete values in the variable of interest or an

end goal of estimating just the extreme values instead of the entire distributions are examples of such cases. The operational interest may many times determine how the clustering should be done.

Further possibilities to identify the subpopulation structure and consequently making the estimates better include making the framework iterative. Comparing the final predicted distribution to the real distribution using cross-validation may let us evaluate how good a particular clustering approach actually is. Iteratively using evolutive optimization algorithms with cross-validation may let us improve the clustering process and help us to find the subpopulation structures. However, these and other theoretical developments are not consider in this study further, but rather next we examine the practicalities related to applying the framework by taking a deeper look to operational interests relating to the passenger weight data that has been the dummy data set in our previous examples.

Chapter 4

Case Example: Passenger Weights

This case example explores the passenger weights and related demographic data that the airline company of consideration has collected. The purpose of this example is to show how the distribution-based subpopulation framework can be applied in practice and what kind of insights it may provide. We mainly apply the most defined and also simplest methods in the framework as they work very well as the data is of good quality.

4.1 Case Overview

The standard weights used in the passenger aviation in Finland are determined by the EASA regulation. [9] These weights, however, seem to be outdated, as it is believed that the population has gained weight since the last measurements, and also the weights have not been especially determined for the airline company of interest. Thus, a weight survey was conducted to determine the new standard weights. The survey included measuring weights of the passengers, and whilst taking the measurement, the persons being measured answered a few questions about themselves and their flights. The survey partly follows the example set by a previous survey conducted by NEA [3].

Using the correct standard weights makes it easier to estimate how much cargo can be loaded and amount of fueling caused by adjusting the amount

of fuel needed for safe aviation. Thus, having an accurate estimate of the total weight of all passengers beforehand is useful. The current regulation [9] requires that it is considered how many male, female and children are occupying the flight. Each of these groups have a standard weight which is used to calculate the estimate for the total weight of all passengers.

On top of just determining the new standard weights as the regulation states, the new survey data raises a few interesting questions and applications to make changes to operations of the airline company of interest, which are shortly described next.

4.1.1 Subpopulation Clustering and Destination Segmentation Problem

Having many specific standard weights for different subpopulations can be beneficial for flight operations. On the other hand, too many standards may be difficult to manage and comprehend. We take a look at this problem of segmenting passengers subpopulations reasonably.

The segmentation to two genders and children, that essentially is the current segmentation, is likely the most meaningful division for the entire population, but further segmentations may be done utilizing other variables such as the traveling class or the destination of the flight. The determination of the most important variables or more general clustering for passenger segmentation, so that the subpopulations would have a common understandable shape and significant differences in the mean values, is desired. Especially, as the destination of the flight is strongly connected to the fuel costs and easily accessible variable for each flight, we should take special attention to see how the segmentation of the flight destinations should be done. Furthermore, the flights are already classified to 4 geographical categories and we can see how usable these are with respect to the passenger weights.

4.1.2 Rare Groups Estimation and Sample Size Reduction Problem

Taking the measurements is laborious process and relatively many measurements are needed for the required statistical accuracy for the new standard weights. This leads to problem of estimating accurately the average weight of a subpopulation of a low number of samples or even with no samples. Also estimating small quantiles, such as, people with more handbag weight than allowed, is interesting for the airline company.

4.2 Overview of the Data, Subpopulations and Distributions

The data set used in this analysis consists of 5090 weight samples collected during years 2017 and 2018. The weighing process was based on voluntary participation at the airport. The measurements represent the customer base of the airline company, other than that the children are excluded. To improve the accuracy, a stratification to genders, geographical regions and seasons was done. As mentioned, the survey follows to example set by an other weight survey shown in [3], which has partly inspired the regulation [9] that is followed when calculating the required sample sizes and other formal aspects about the reporting and conduction of the survey. Note especially that the gender ratio is regulated so it does not reflect the true gender ratio for the passenger, but the other variables (besides the season variable) should reflect the reality reasonably well. Note also that this study only contains the first 5090 samples collected for standard weight revision and we do not consider here questions related to possible biases in the survey or other problems. The calculations may also be simplified here to what the regulations requires for the actual survey study that is not this thesis.

Each weight sample is the sum of a passenger pap weight and handbag weight. The pap weight, that is human body weight with clothes, was measured using a larger scale and the handbag weight was weighed on a smaller scale. All the total weight measurements are shown in Figure 4.1 and by pap and handbag weights in Figure 4.2. The participants also answered a collection of questions while being measured. The survey device also automatically recorded the time of measurement and the flight number. The season and region are

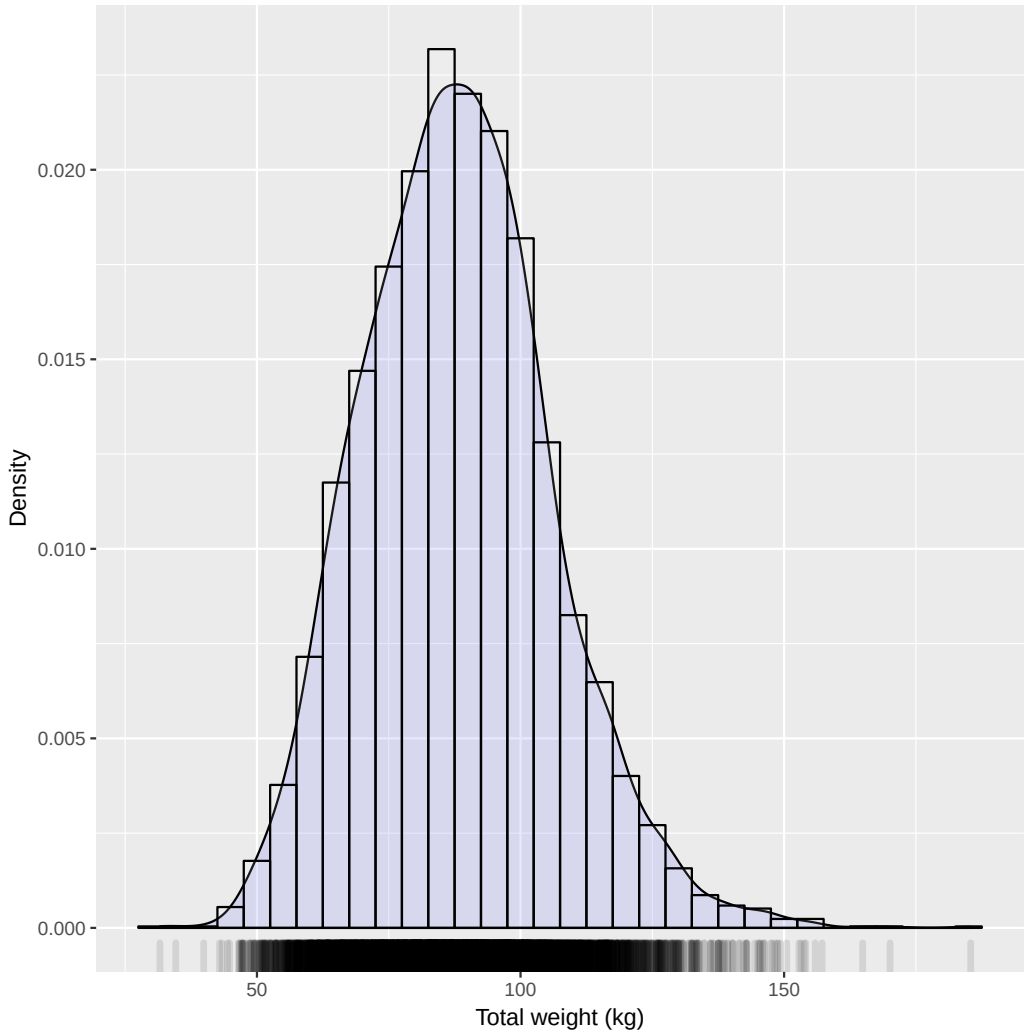


Figure 4.1: All total weight measurement samples and the density estimation for them.

not directly asked but can be determined from the data naturally. The answers to survey questions are shown in Figure 4.3 with respect to number of answers to each question and in Figure 4.4 the average weight for each answer subpopulation. The questions should be rather unambiguous from the answer, other than the language, which refers to the language the participant chose to answer the questions with rather than the actual nationality.

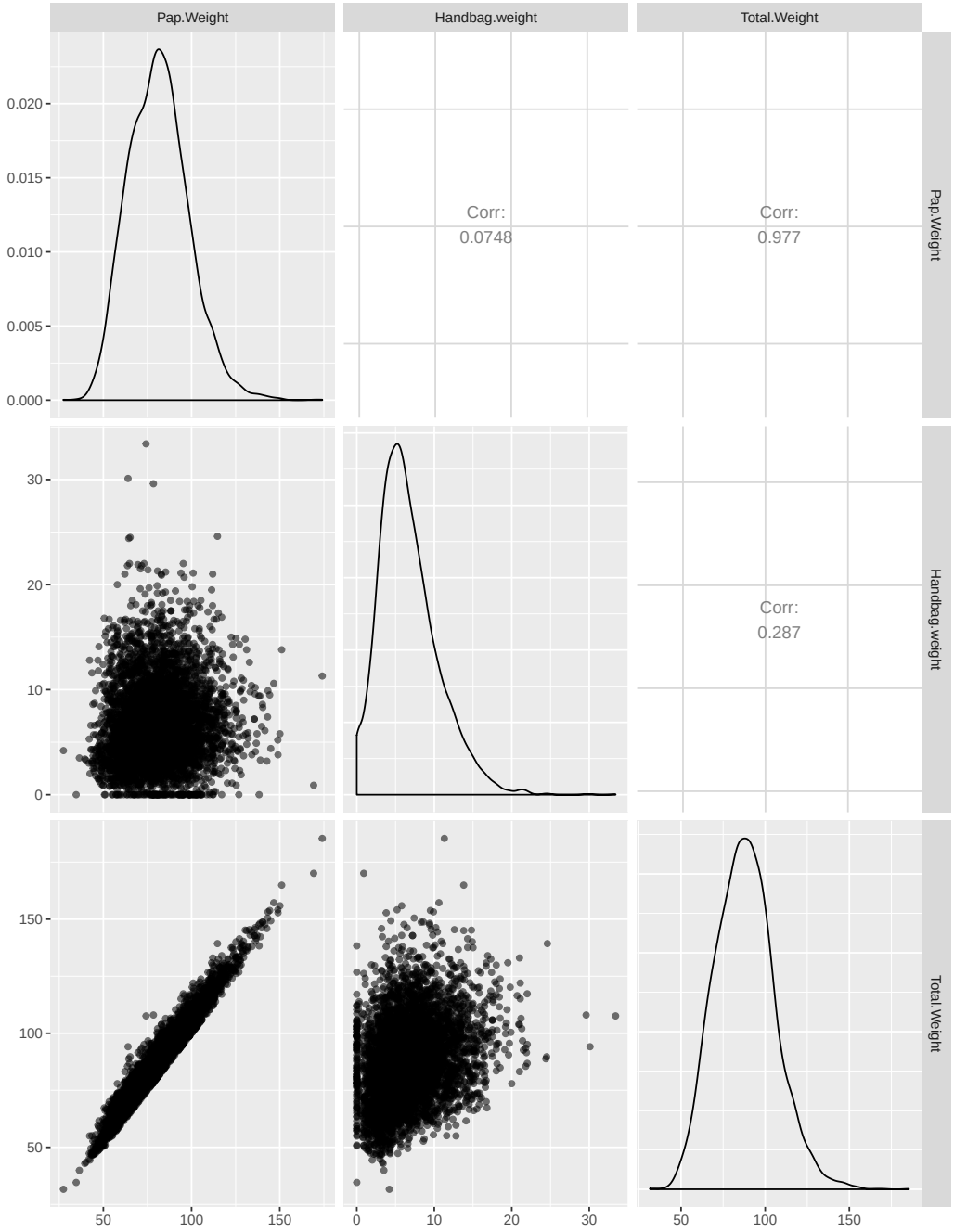


Figure 4.2: All weight measurements: total, pap and handbag weights and their relations.

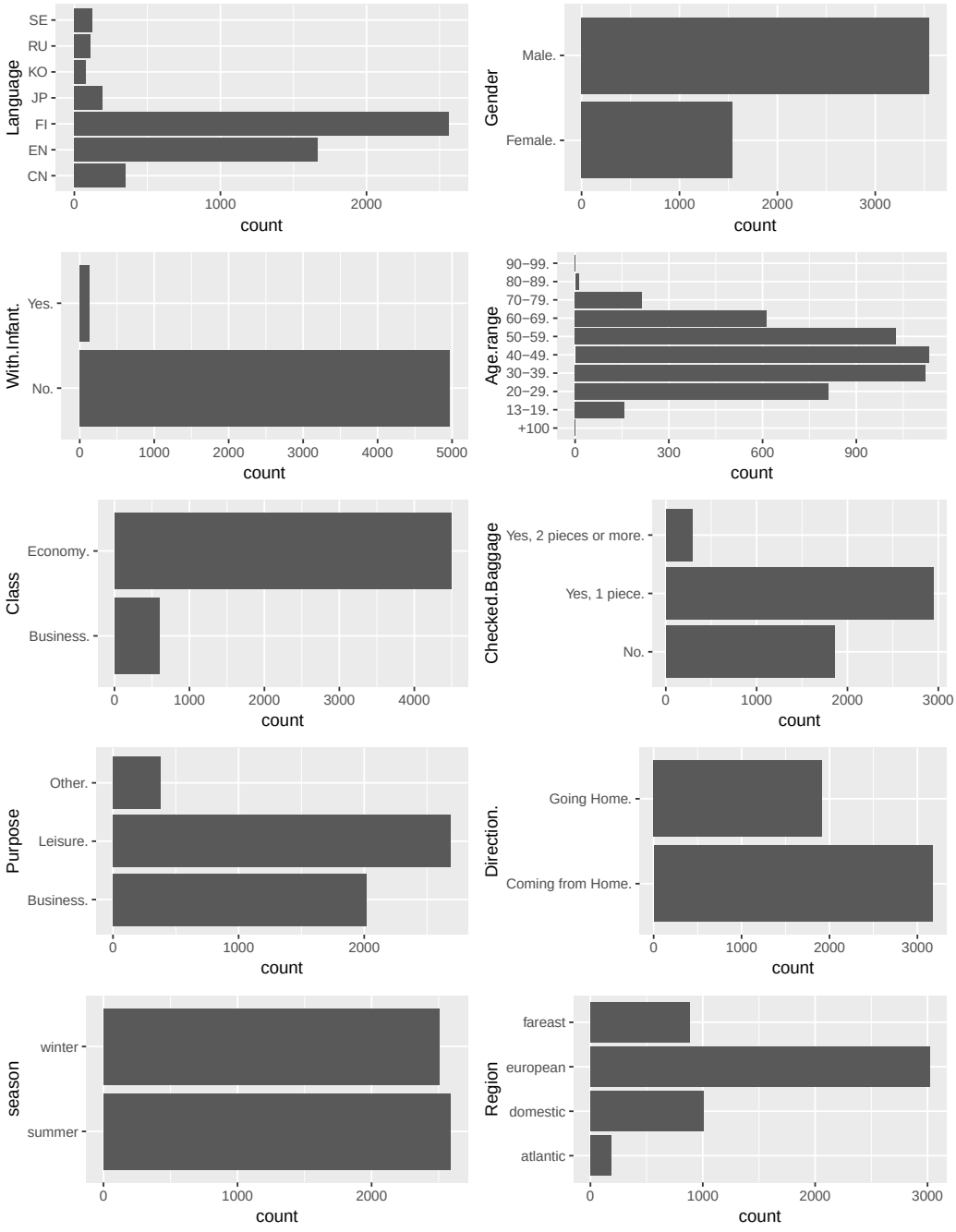


Figure 4.3: The counts of answers to each question in the survey.

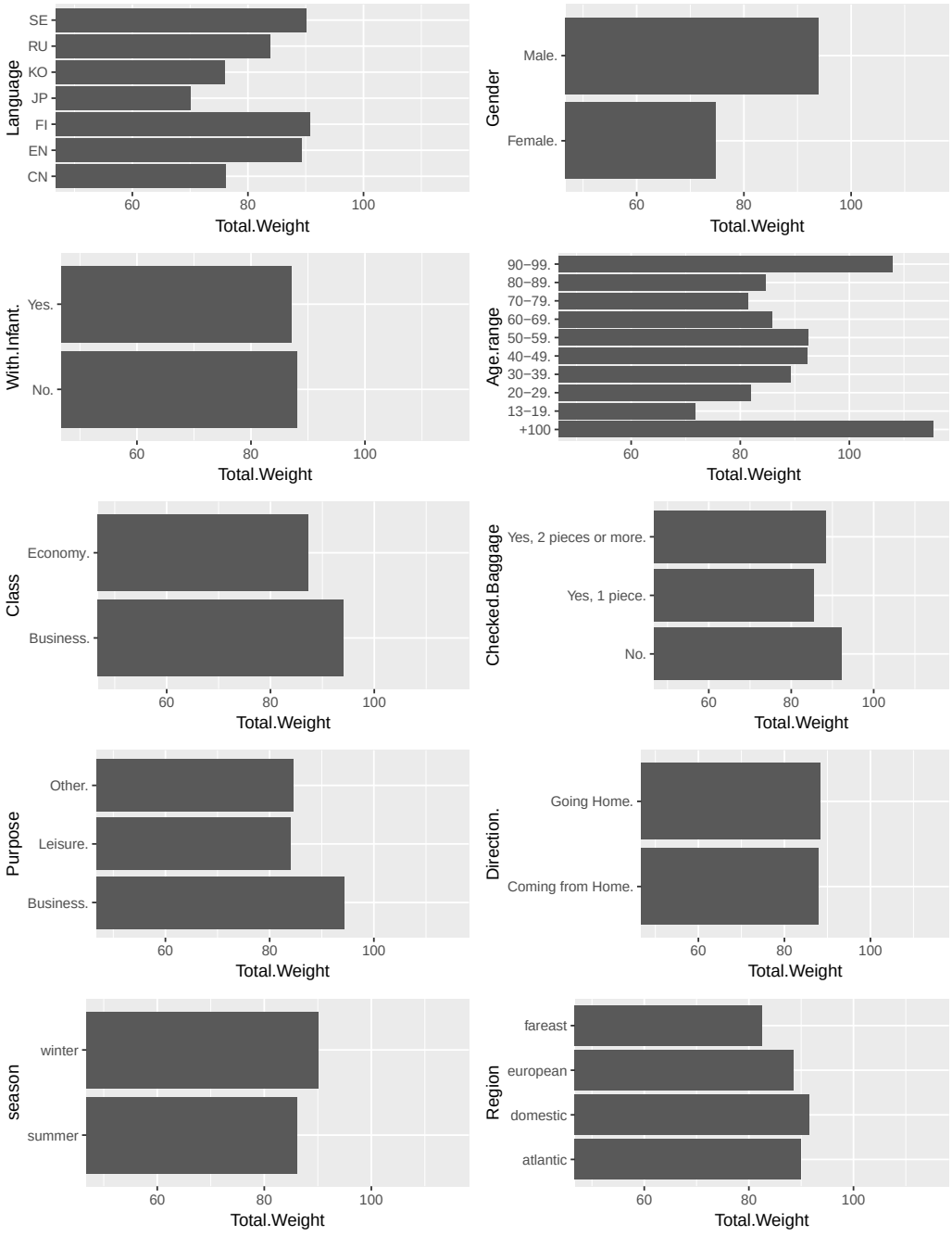


Figure 4.4: The average weights for each subpopulation in the survey.

To understand how each categorical variable affects the total weight, we form a linear regression model. The model shows the effect each category has to the total weight and the statistical significance. The results of the model are shown in Table 4.1. These results are further analysed and used in the categorical clustering in the next section.

Table 4.1: The linear regression model for all categorical variables in the survey data to explain the total weight variable

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	95.7366	14.1306	6.775	1.38e-11	***
LanguageEN	8.8527	0.8863	9.989	< 2e-16	***
LanguageFI	11.3508	0.9063	12.524	< 2e-16	***
LanguageJP	-6.2435	1.2683	-4.923	8.80e-07	***
LanguageKO	-2.6183	1.7498	-1.496	0.134627	
LanguageRU	7.6987	1.5625	4.927	8.60e-07	***
LanguageSE	10.5437	1.5131	6.968	3.62e-12	***
GenderMale.	17.2688	0.4497	38.403	< 2e-16	***
With.Infant.Yes.	-0.6296	1.2741	-0.494	0.621251	
Age.range13-19.	-37.4101	14.0654	-2.660	0.007845	**
Age.range20-29.	-26.3673	14.0319	-1.879	0.060289	.
Age.range30-39.	-22.5332	14.0300	-1.606	0.108321	
Age.range40-49.	-19.4274	14.0297	-1.385	0.166194	
Age.range50-59.	-18.8266	14.0291	-1.342	0.179667	
Age.range60-69.	-21.6572	14.0325	-1.543	0.122806	
Age.range70-79.	-25.1047	14.0528	-1.786	0.074086	.
Age.range80-89.	-28.5010	14.6444	-1.946	0.051686	.
Age.range90-99.	-9.7685	17.1810	-0.569	0.569678	
ClassEconomy.	-1.6140	0.6484	-2.489	0.012838	*
Checked.BaggageYes, 1 piece.	-1.8259	0.4610	-3.961	7.57e-05	***
Checked.BaggageYes, 2 pieces or more.	-0.1355	0.9174	-0.148	0.882567	
PurposeLeisure.	-1.7734	0.4829	-3.672	0.000243	***
PurposeOther.	-0.9343	0.8403	-1.112	0.266238	
Direction.Going Home.	1.4315	0.4479	3.196	0.001402	**
seasonwinter	1.2306	0.4187	2.939	0.003307	**
Regiondomestic	-4.2497	1.1697	-3.633	0.000283	***
Regioneuropean	-3.8430	1.0880	-3.532	0.000416	***
Regionfareast	-4.6419	1.1817	-3.928	8.68e-05	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

4.3 Subpopulation Clustering and Destination Segmentation

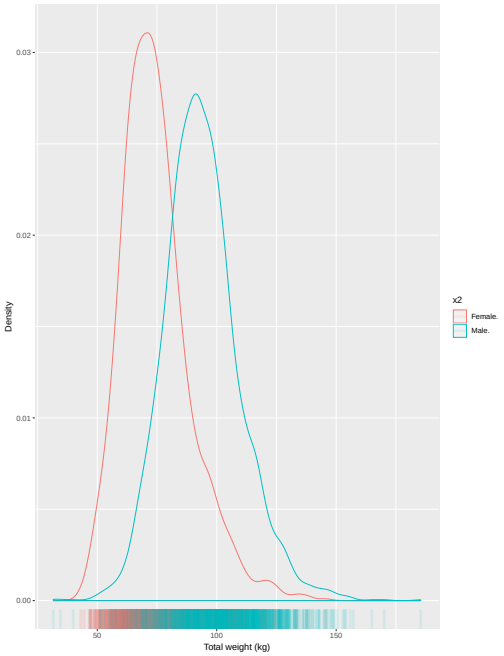
We do the clustering to subpopulations using the categorical and hierarchical clustering. The survey data variables are already categorized other than time, pap and handbag weight. Time is categorized post-measuring to two seasons, and pap and handbag weight are not really useful, because we need to measure the passengers to access these, which cannot be done in regular traveling. Also, the number of different flights is relatively large, and categorized to 4 different geographical area, and this variable is further analysed separately later in the destination segmentation section.

The segmentation should be applicable in daily flights, have large enough differences in the means, have as few groups as possible, and it should be easy to assign a passenger to the right segment before the flight. We try to conclude a segmentation that matches these criteria by the end of this section. To start with we use the linear regression model and the variable counts and average tables to choose how the categorical clustering is done. We exam the most prominent variables and their combinations for making the subpopulation distributions.

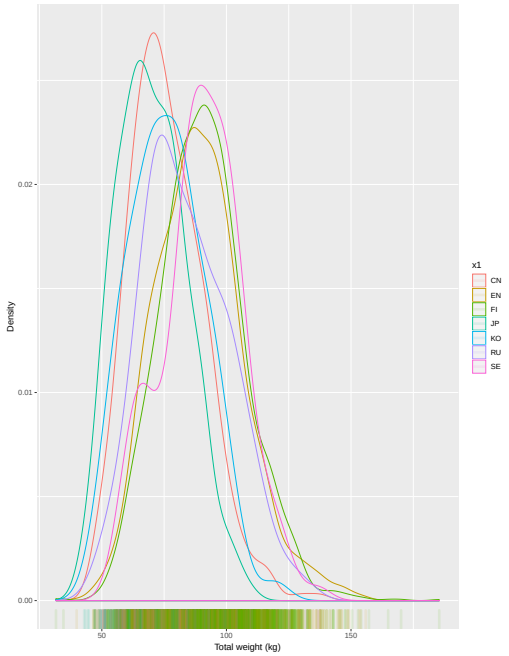
4.3.1 Results: Subpopulation Clustering of the Weight Data

We consider the categorical group clustering of the data using the existing categorical variables. The most important variable seems to be the gender. These subpopulations have a clear difference in the average weights and variations also. The gender distributions are shown in Figure 4.5(a). The overall shape seems otherwise very similar, but the lower tail is steeper for the females. Both of the distributions have a longer upper than lower tail and resemble the lognormal distribution.

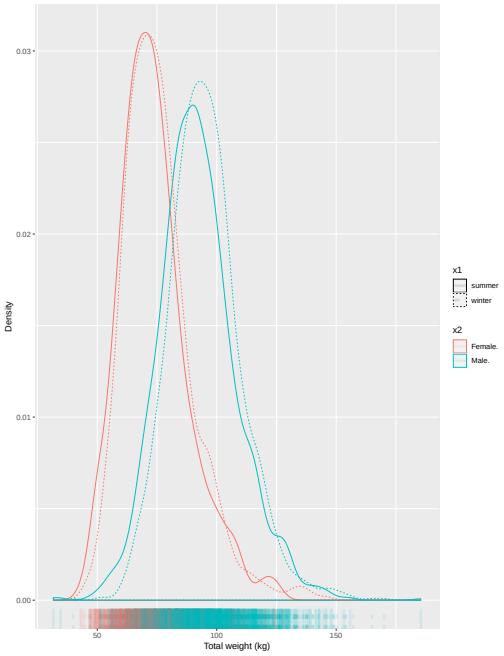
The age seems to have a strong effect but it has a low statistical significance other than the difference between teenagers and adults, for which the teenagers naturally weight less. This kind of segmentation is though already made, as the children have a separate standard weight value in the regulation.



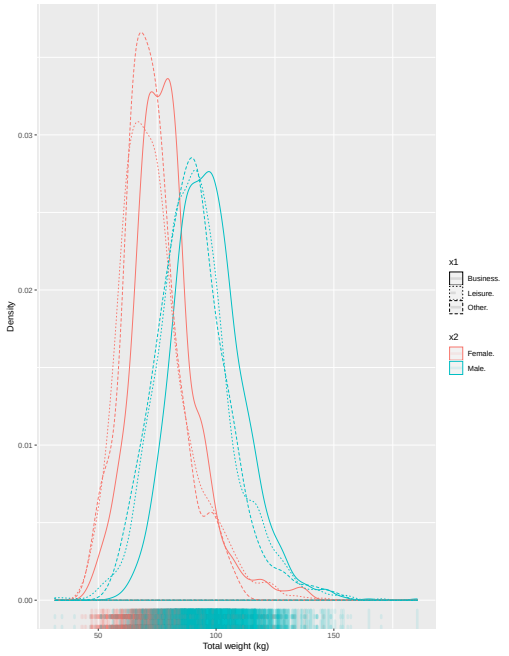
(a) Distributions by the gender subpopulations.



(b) Distributions by the language subpopulations.



(c) Distributions by the season-gender subpopulations.



(d) Distributions by the purpose-gender subpopulations.

Figure 4.5: The weight distributions by different subpopulations based on various categorical variables.

The language and region both have a clear effect. Since these are correlated heavily, we only consider the language in Figure 4.5(b). Resulting shapes are similar to the gender distributions. The Finnish, English and Swedish subpopulations seem to be very similar in shape and have similar lower tails than the male subpopulation. Chinese, Japanese, Korean subpopulation have a lower tail like the female subpopulation. Russian is something in between what comes to the lower tail. When categorization is done with respect to both the language and gender we get possibly the most similarly shaped, but also distinct, collection of subpopulations of relatively many clusters and for which we have large variations in the means. This clustering is visualized in Figure 4.6.

Examples of other clusterings include the gender and season clustering seen in Figure 4.5(c). The shapes are well-defined but differences in the means are too small to be commercially usable. The purpose of flight, seen in Figure 4.5(d) with the gender variable, is also a rather significant variable and subpopulations have considerably different means using this variable too, but the overall shapes are not that well in unison as in the gender-language case. However, pretty much all the variables manage to generate relatively similar shapes at least as long as the gender is used in categorization.

We also examine how well the hierarchical clustering can find these subpopulation structures. Using all the categorical variables to predict the total weight using the Gower distance with equal weights we get the clustering displayed in Figure 4.7. There does not seems to be a clear cut of value. This could indicate that there are no meaningful clusters, as if the data were random and without any patterns, but in our case it likely indicates that there actually multiple meaningful ways to do the clustering. The subpopulation distibutions are shown in Figure 4.8 for ten different cut-off values. A common shaped structure for the distributions seems to exist. However, distribution are mainly stacked to two pile even though there were no very clear cut-off value to two clusters.

We will further consider the hierarchical clustering by selecting important variables using domain knowledge next, but with no domain knowledge the best clustering seems to be the division to the genders that is already in use in aviation industry, and if we wish to expand this segmentation, either the area or language with the gender variable would provide the most meaningful and predictable clustering. Also variables, such as the age or class, could be interesting to consider further if we had some specific operational question relating to these in mind.

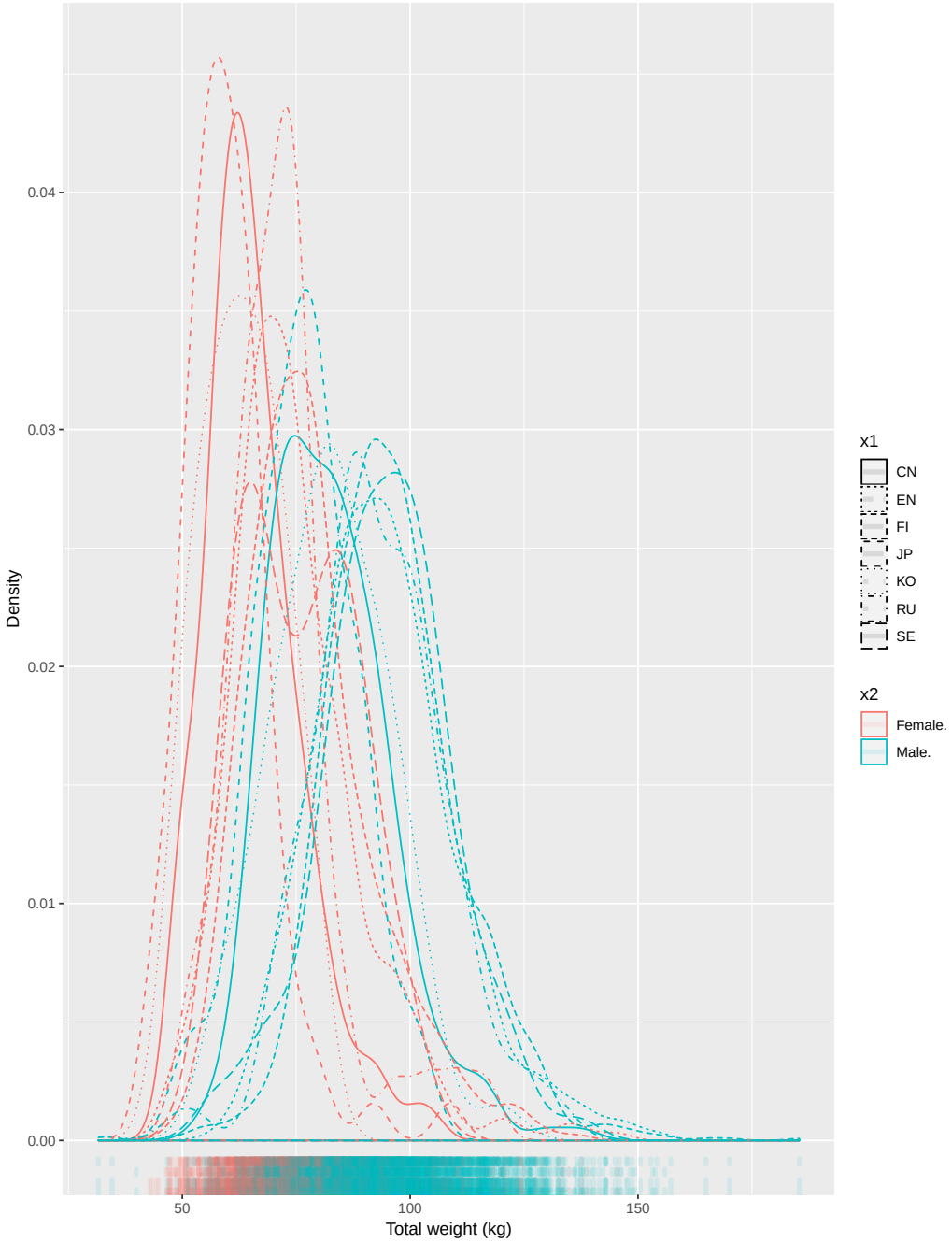


Figure 4.6: The weight distributions by the language-gender subpopulations.

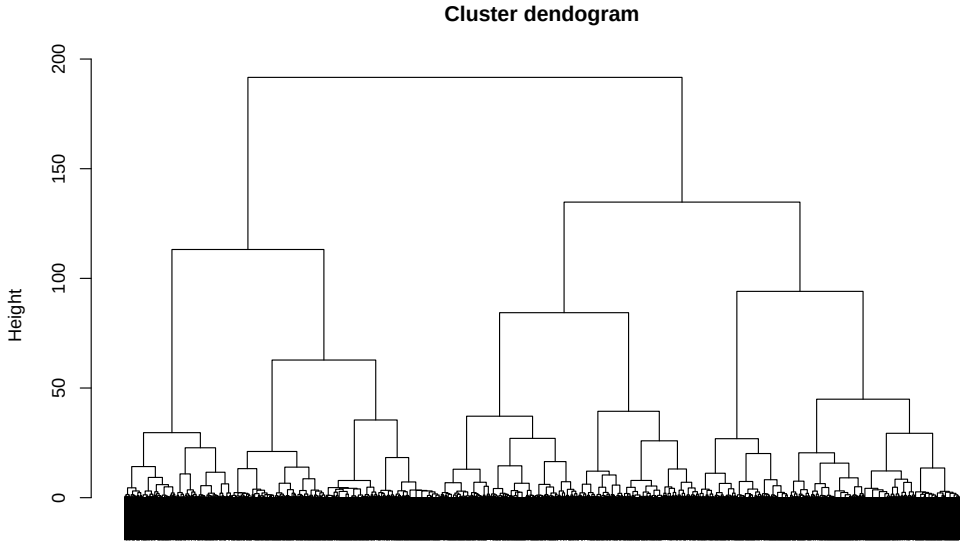


Figure 4.7: Hierarchical clustering dendogram using the Gower distance for all the categorical variables in the weight data.

4.3.2 Results: Segmentation of Destinations

We reconsider how the variable flight number should be aggregated to smaller amount of categories. The flights are listed by the destination city in Figure 4.9, and in the world map in Figure 4.10. Each destination city also consists of different flight numbers that are in the data, but these are not used in the analysis here. The current categorizations include the aggregation to nations and the regional division to Domestic, European, Atlantic and Fareast flights.

The clustering of destination cities is done using the hierarchical clustering choosing the Gower distance so that it emphasizes the location of destinations in the world, the average total weights in that area and the number of flights. The clustering dendogram, when using even weights for location and average weight in the area, can be seen in Figure 4.11, the number of flights getting weighted automatically as the number of samples is proportional to the flights per year. The dendogram suggests a clear cut-off value to 2 clusters. Also 3 clusters could be considered if operational domain knowledge supports this.

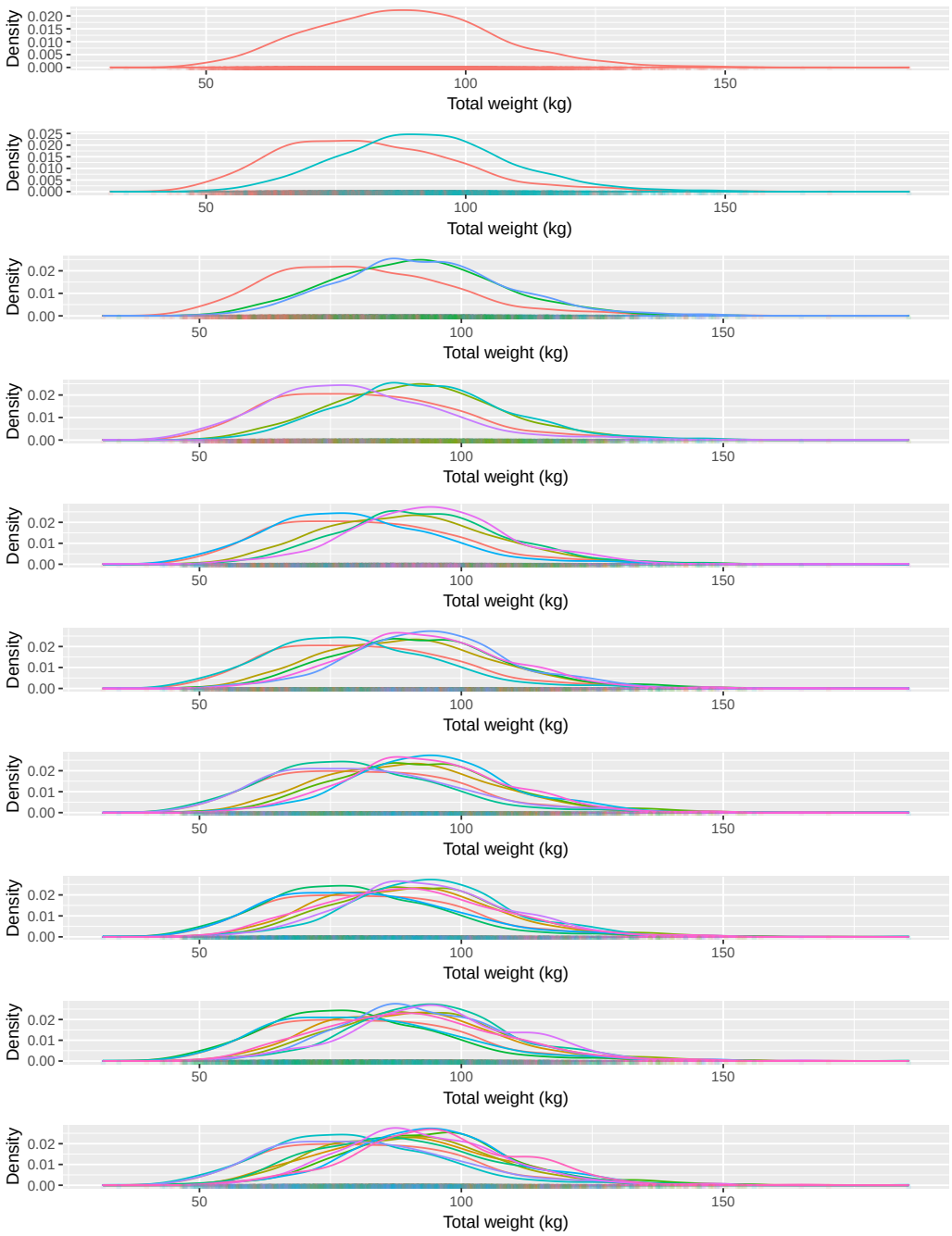


Figure 4.8: The distributions using the hierarchical clustering with different cut-off values for Dendogram 4.7.

To consider the clustering in practice with domain knowledge insights, let

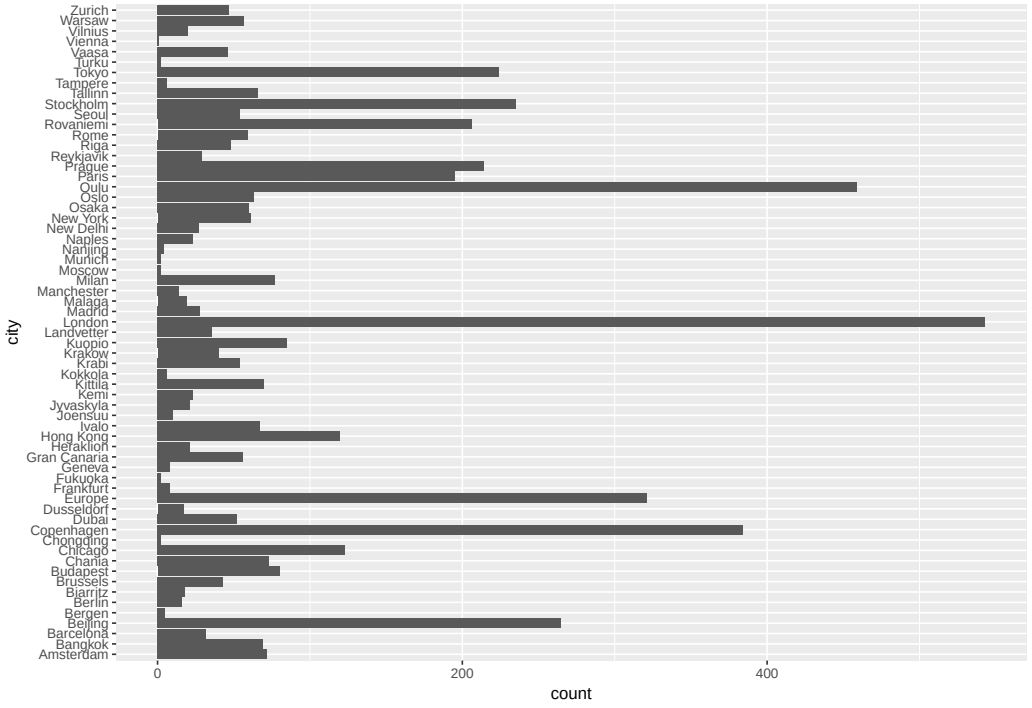


Figure 4.9: The amounts of measurements per destination city.

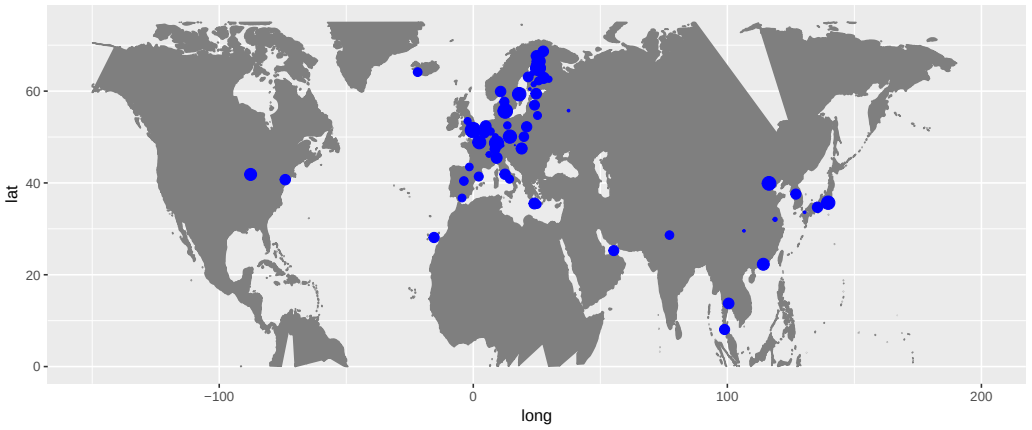


Figure 4.10: The destination cities measured on the world map, the number of flights being logarithmically scaled.

us examine the clusters on the world map. Different cut-off values for Dendogram 4.11 are shown in the world map in Figure 4.12. For comparison it is also shown how the clustering is done if only the local average weight is considered in Figure 4.14 and if only the location is considered in Figure

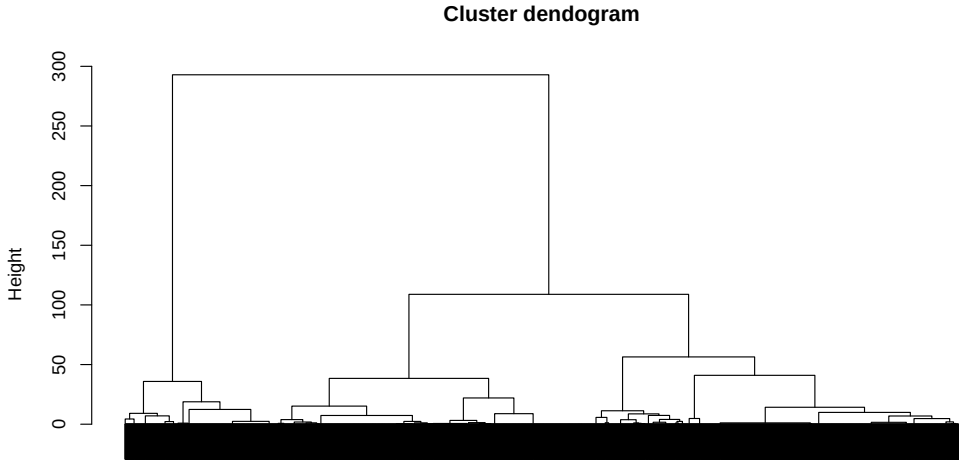


Figure 4.11: Dendogram using the Gower distance with equal weights for the average area weight and the location.

4.13. The Figure 4.12 shows two easily understandable segmentations. The first one being Fareast and rest of the world. The second one being same as the previous but the Nordic countries are separated to their own cluster. Figure 4.13 and 4.12 provide further insight about a practical clustering. The indications are similar to the analyses with 4.12, but for instance, we could expand the Nordic cluster with Central Europe destinations in case of using 3 clusters. We can also detect some individual particularly light or heavy destinations in certain geographic areas, such as the two purple destinations in 4.12 or the Atlantic destinations.

When making a decisions about the standard weights and subpopulations to which they are put into practice, we must note that having too many segments can be unpractical to understand and it also requires more weight samples for required accuracy of statistics. Also different subpopulations are easier to separate in daily operations than the others. For instance, we do not have direct access to the language the passenger speaks but we do know their flying class and gender. Without using further domain knowledge, such as ticket prices, new operational changes and ignoring details like that the fuel consumption is not entirely linear to traveling distance which is assumed in this analysis, we could suggest segmenting the world to 2 or 3 segments

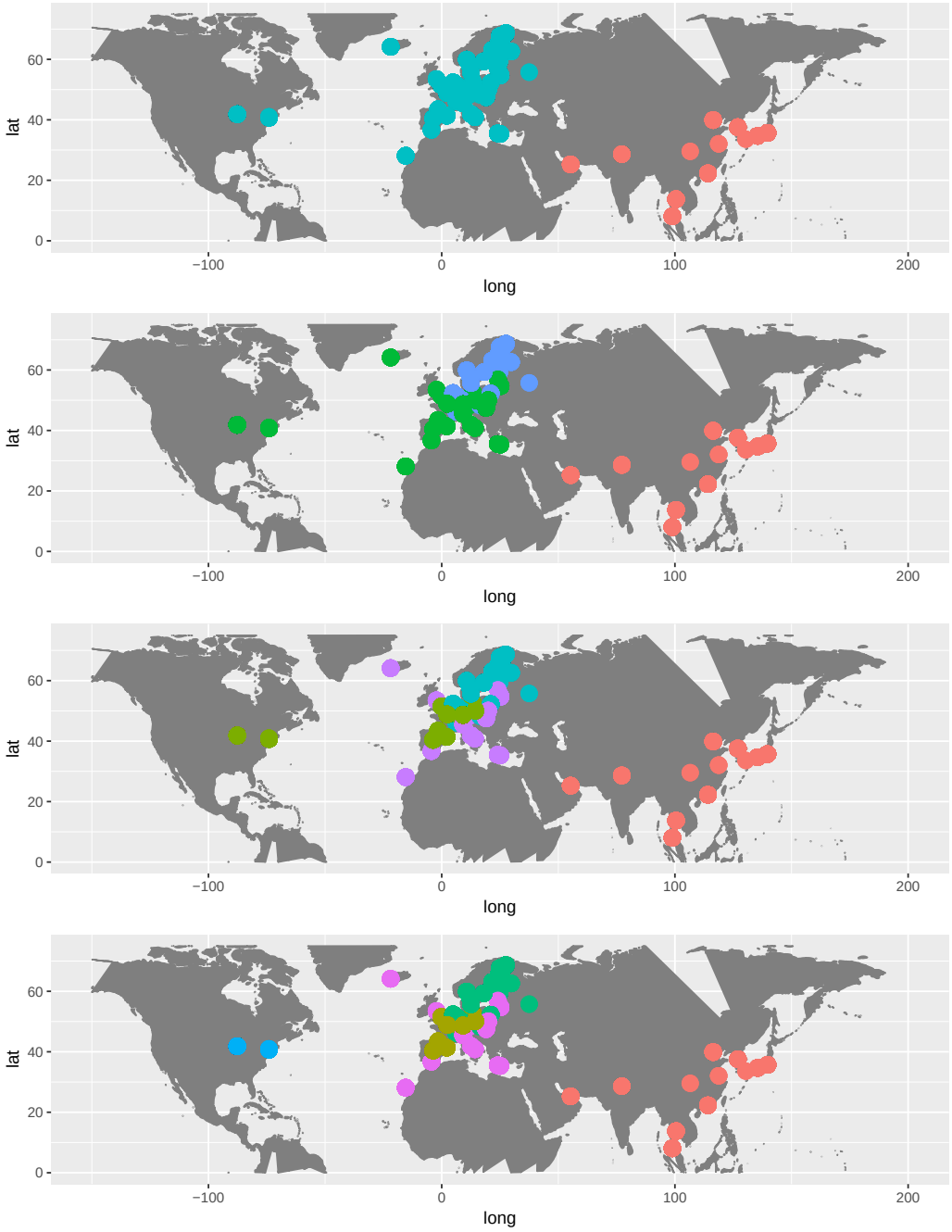


Figure 4.12: The clusters on the world map, equal weighting used for the average area weight and the location. See the related dendogram shown in 4.11.

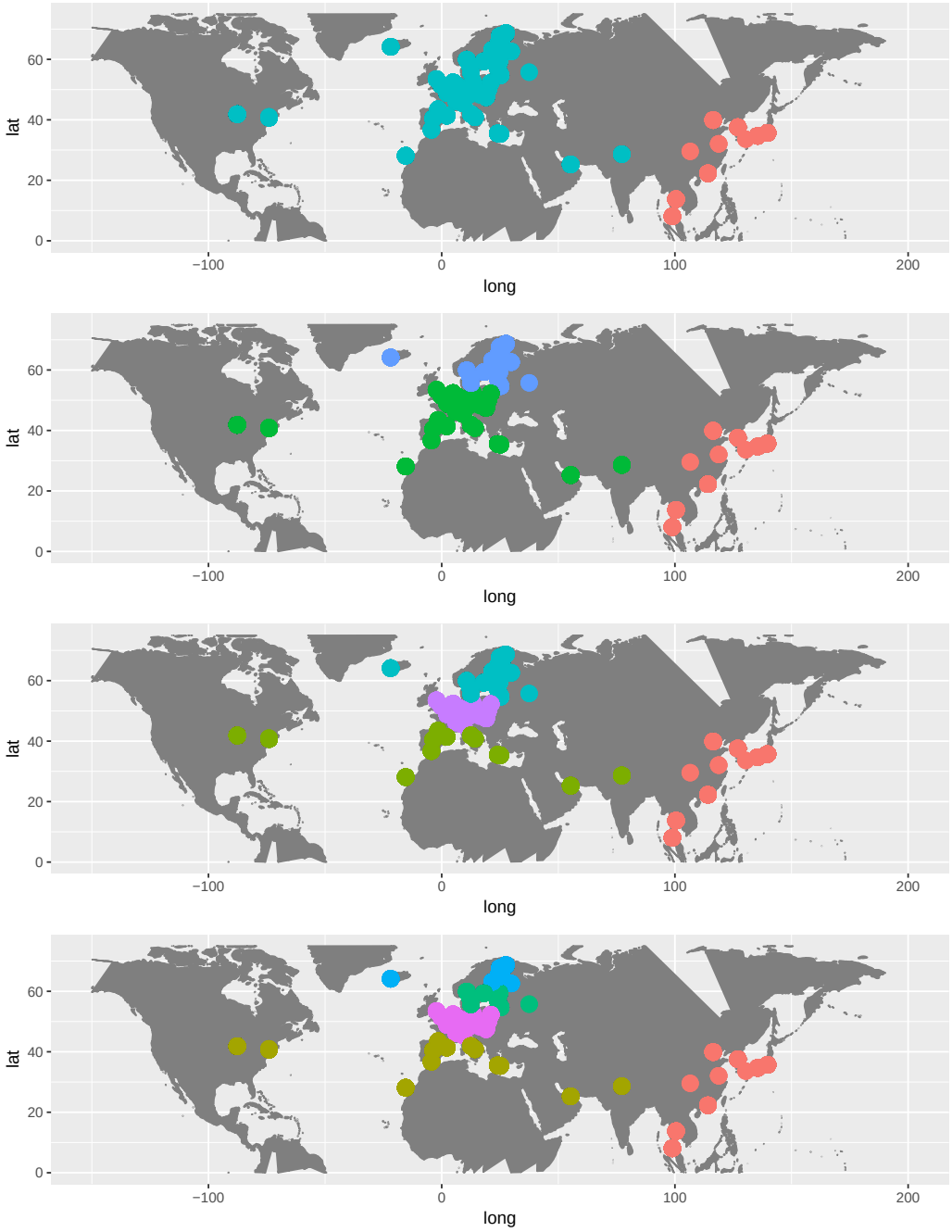


Figure 4.13: The clusters on the world map using only the location as a clustering criterion.

that have been discussed here.

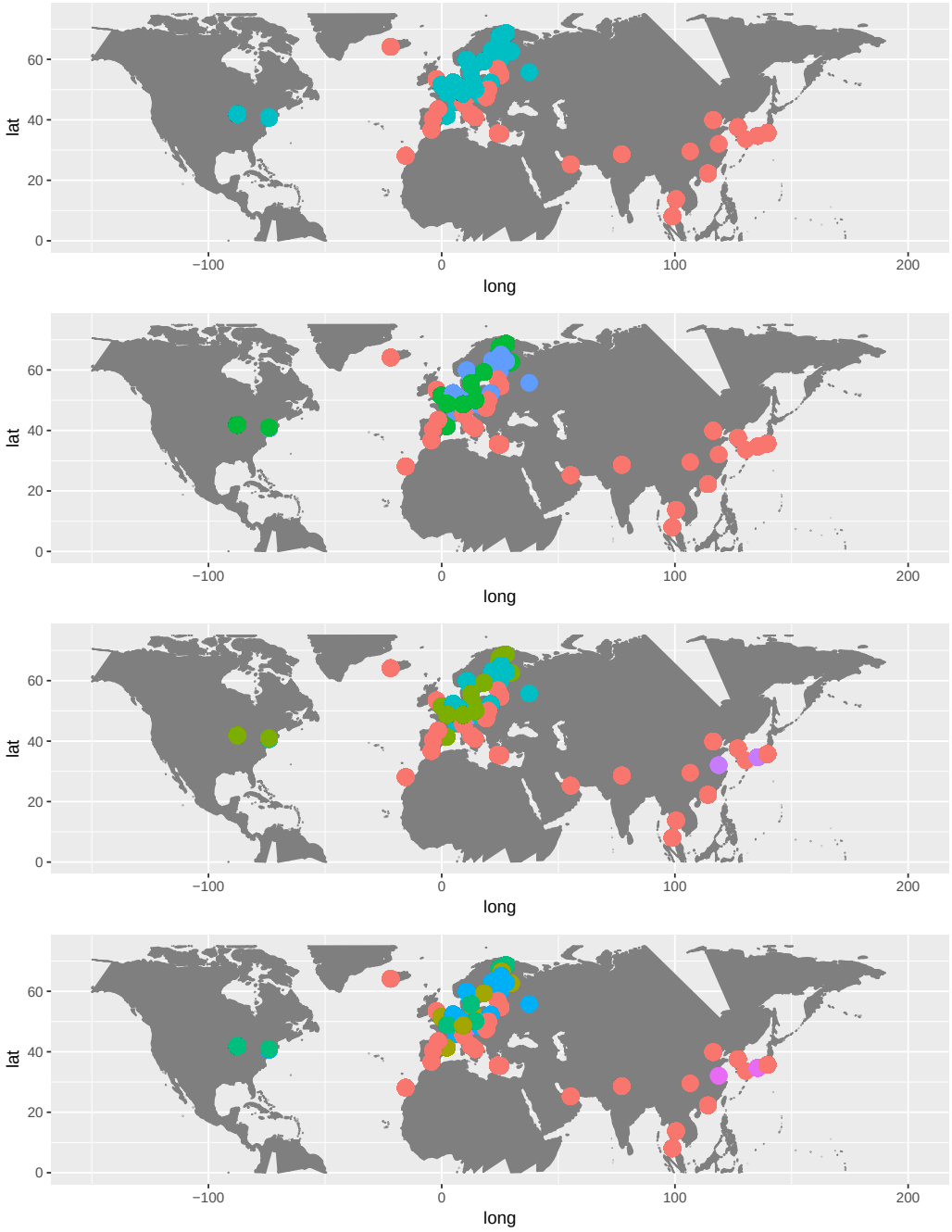


Figure 4.14: The clusters on the world map using only the average area weight as a clustering criterion.

4.4 Estimations for Small Groups and Rare Events

In this section we construct subpopulation distribution estimations based on different amounts of samples and compare them to smooth fits references. We also compare the tail behavior and extreme value estimates.

4.4.1 Results: Subpopulation Estimation

We compare how the use of shape identification methods improves the distribution shape predictions for the weight data subpopulations. To evaluate the quality of a method we compare it to the true underlying distribution. Here we examine the subpopulation of Finnish males for which we have defined the shape using all the samples in that group and doing a kernel density estimation fit, which we consider to be the true underlying distribution shape. We then bootstrap random samples from this subpopulation and apply our methods to do the validation. In practical applications, these methods should be used for some population from which we have only a little samples and we want more accuracy for the distribution shape, but this would not enable us to do desirable validation.

The example subpopulation clustering is done using the categorical clustering using the gender and language variables, see Figure 4.6. The location-scaling averaging method variations are being used. In Figure 4.15 we compare the location-scale method, smooth fit and the true underlying distribution with different number of samples. The example is a presentable illustration how the method usually works. Occasionally, the smooth fit may be even better, but in general the location-scale method produces better results for this kind of data, especially, if we wish to have a good understanding of the distribution shape, but even the raw numerical error is smaller for location-scale method than for the smooth fit in general. It should be noted that the subpopulation of interest is part of the subpopulations used in the algorithm, which slightly improves the results, but overall the method works similarly even if the subpopulation of interest is excluded from the data.

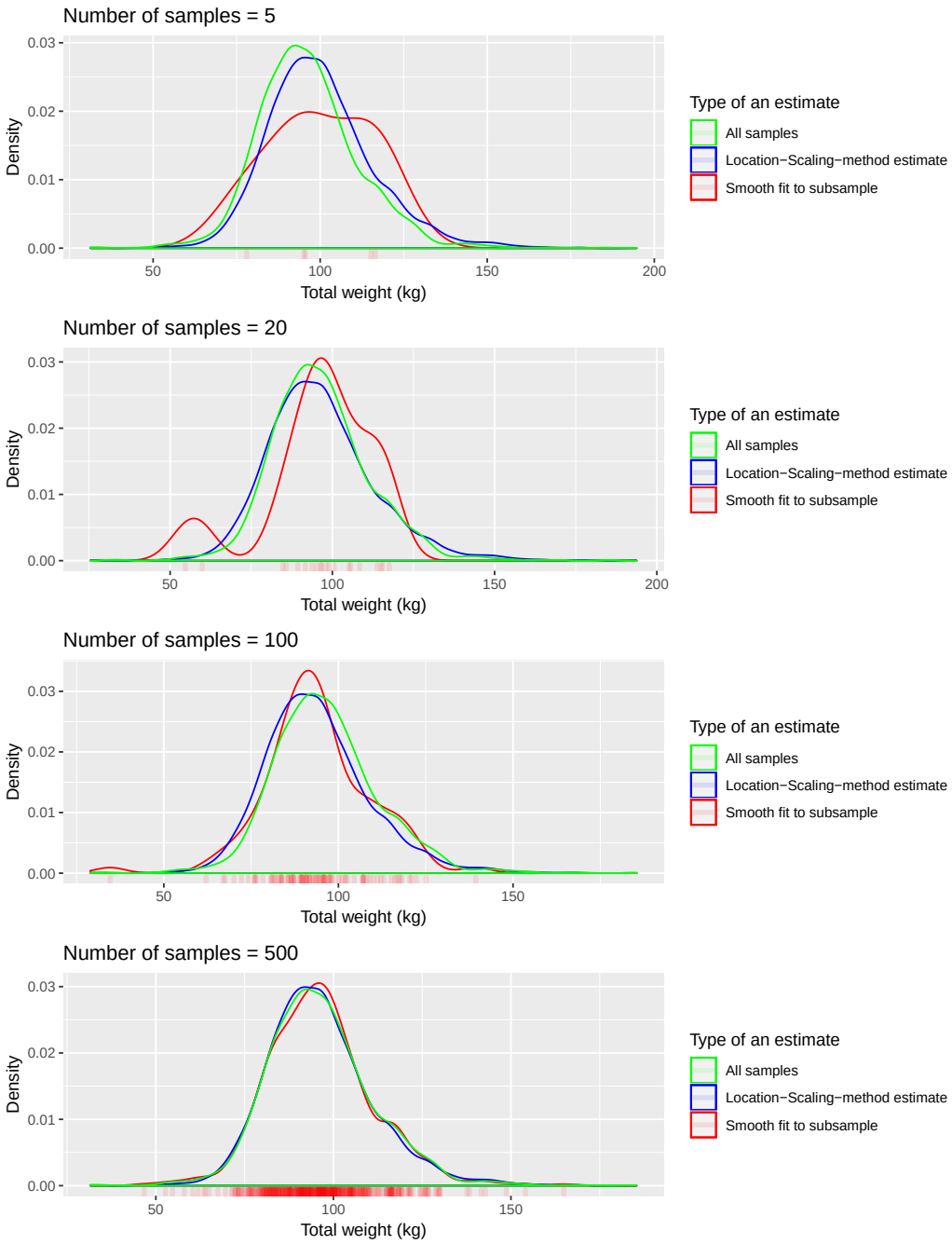


Figure 4.15: The location-scale method estimates compared to the smooth fit and the true distribution using various sample sizes.

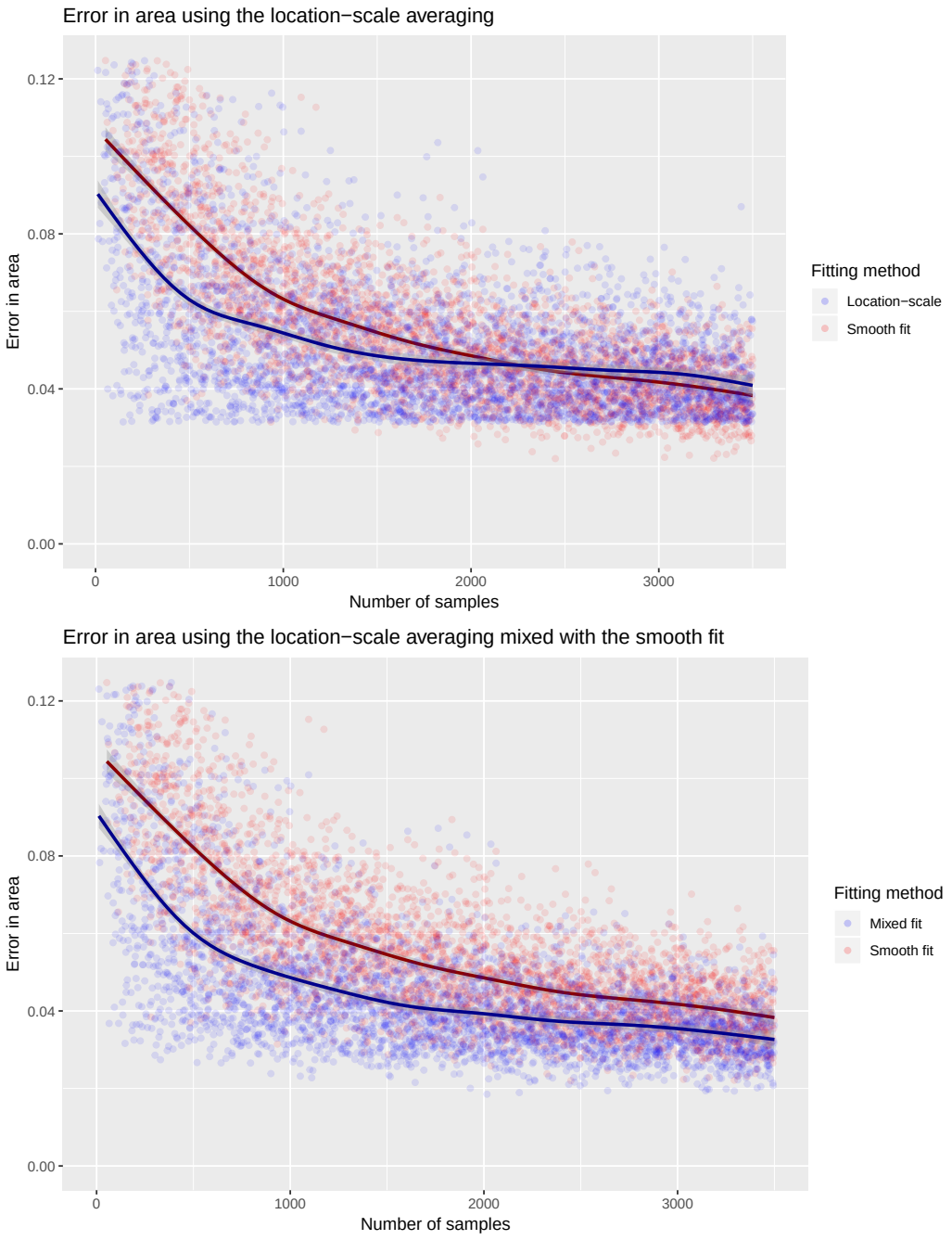


Figure 4.16: The error to the distribution of total weights of all Finnish males with different bootstrapped sample sizes for which the location-scale method, smooth fit and a method that mixes these two are used. The average errors visualized with the trend lines.

The location-scale method provides a good understanding of distribution shape for low sample size cases. As the sample size increases the smooth fit also starts to catch the true shape. It must be particularly be noticed that even if the average of the location-scale fit is not aligned with the true fit, the shape is very similar, unlike the smooth fit.

The goodness of the fit can also be accessed numerically by comparing the area differences between the true fit, as it is shown in Figure 4.16. The samples have been bootstrapped from all the samples in the subpopulation and the error area is calculated using the earth mover's distance. On average, the location-scale method produces smaller error than the smooth fit when the number of sample is small. However, when sample size increases the smooth fit becomes more accurate and should be trusted more. From the upper plot it can be seen that the error of the smooth fit slowly approaches 0 as the sample size increases. The location-scale has however, a limit that it will reach if the sample size is grown which is around 0.03 which it occasionally reaches even with a low sample size. The smooth fit and the location-scale method can also be mixed as explained in the previous chapter. The results of this are shown in the lower plot. The mixture attempts to pick the strengths of the both approaches and seems to be successful.

4.4.2 Results: Rare Event Estimation

Here we take a deeper look at the estimate distribution by looking at their tail behavior. Again we consider the subpopulation of Finnish males similarly to the previous analysis 4.15. Figure 4.17 shows fits to samples that are less than 70 kg and Figure 4.18 to samples that are more than 120 kg.

The absence of points makes it impossible to give any smooth fit estimates in the lower sample cases or at least undesired spikes are more prevalent. The latter problem could be partly solved with a larger bandwidth in the kernel fitting method, but still the location-scale averaging gives a more realistic picture in most of situations in 4.17 and 4.18 compared to the smooth fit to the subsample, which does happen most of the time when this kind of comparison is made. Note that we could also view the actual tails of distributions in the previous analysis, but the analysis here emphasizes even more the problems that the absence of samples causes.

The errors for distribution estimates for the extreme values are considered in

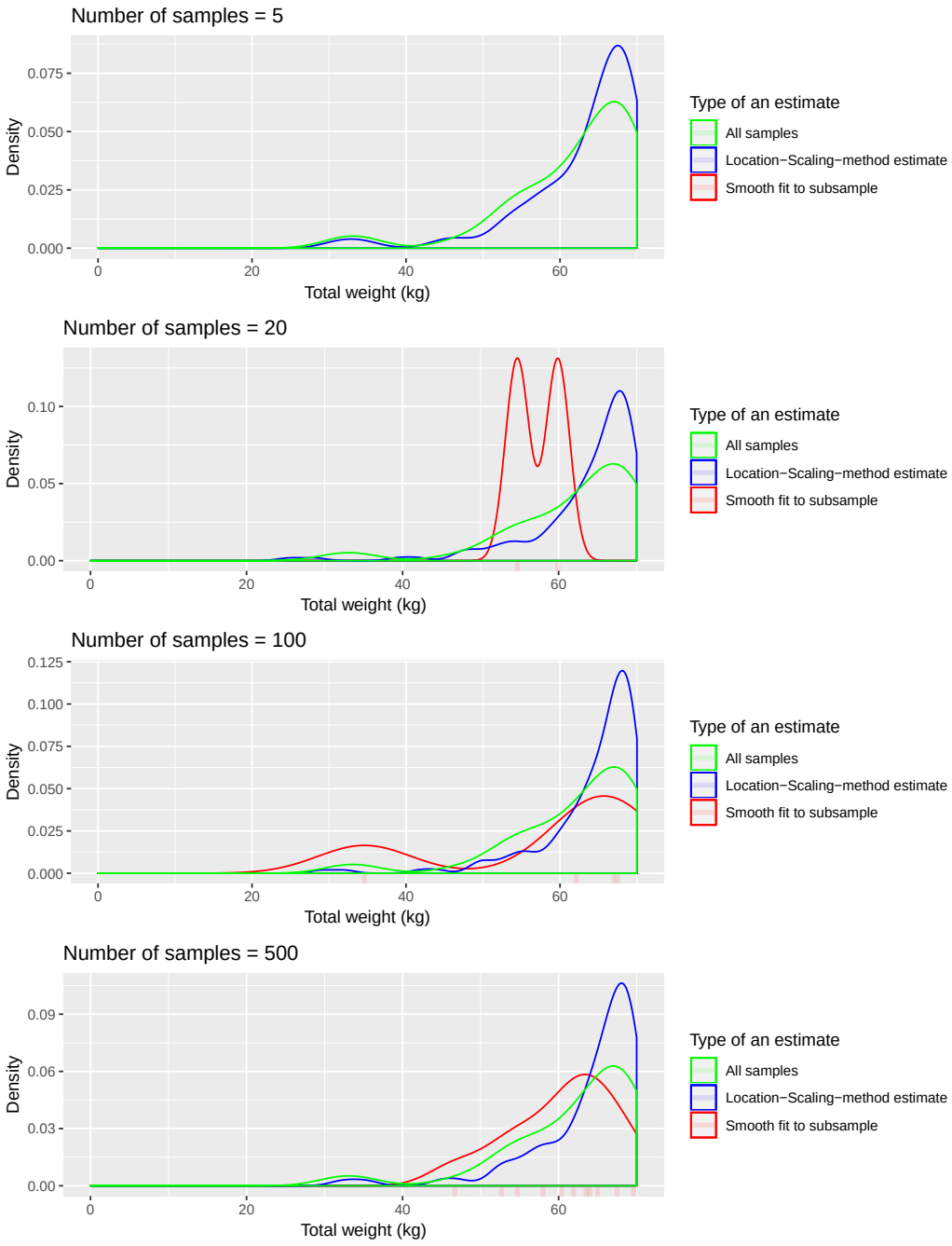


Figure 4.17: The fits of the methods to the samples from the lower tail.

Figure 4.19 in similar manner to 4.16. Again the location-scale averaging is a stronger model on average when the number of samples is not very large.

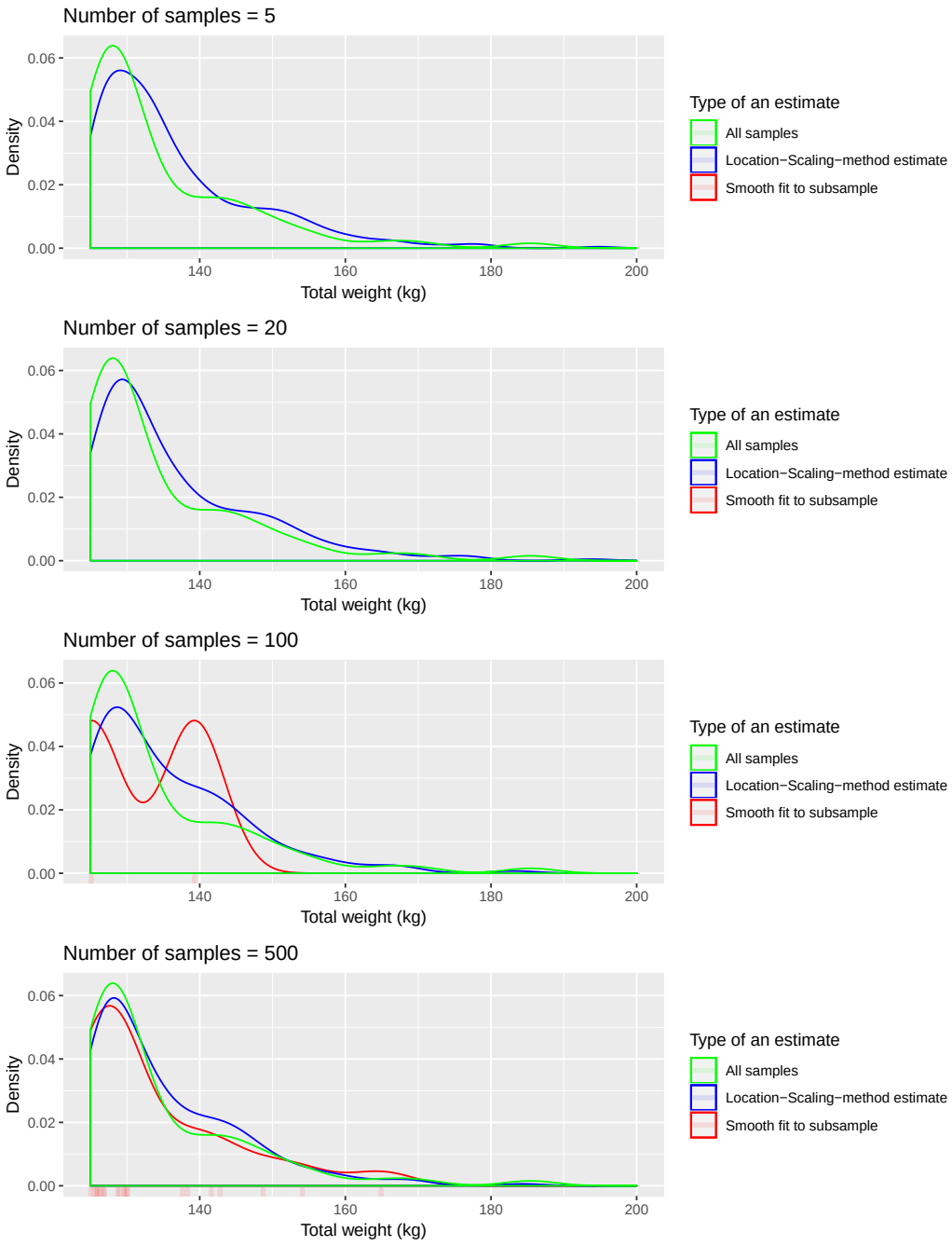


Figure 4.18: The fits of the methods to the samples from the upper tail.

It seems that the location-scale averaging makes much less frequently major estimation errors that the trivial smooth fit.

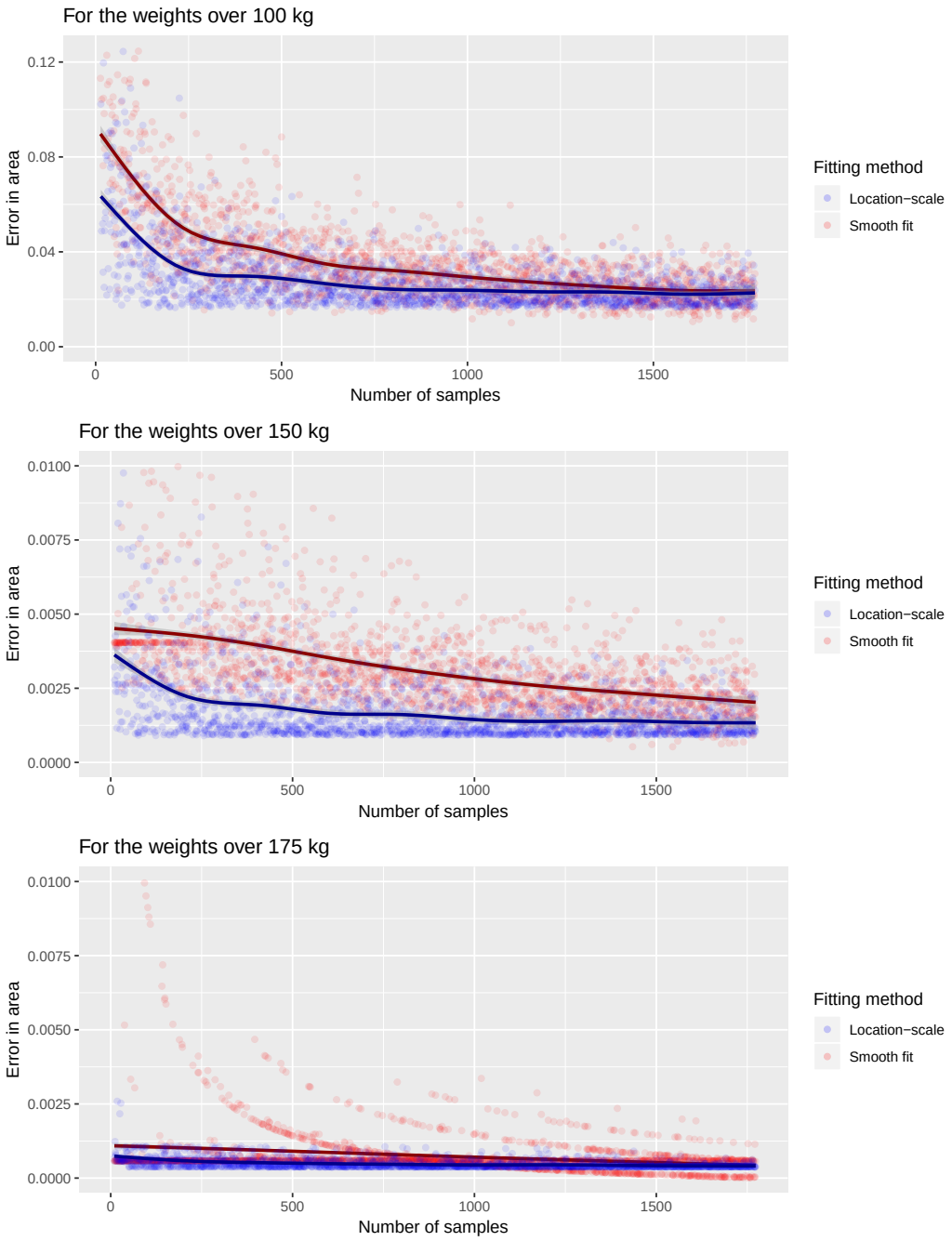


Figure 4.19: The errors for the methods, when the error is measured for the fits to the samples above a selected threshold limit. The samples are bootstrapped from the subpopulation of Finnish males similarly to 4.16.

Consequently, the improved estimates by the framework are applicable for more accurate extreme value estimations. This can be used, for instance, to predict the amount of people with handbag weight more than some threshold limit in a destination for which we have only a few samples. Other examples include the analysis what will happen if a new handbag policy is enforced that mainly affects a small subpopulation or we would like pioneer something new like reserving multiple seats to heavy passengers, but the actual implication for the airline company are left uncovered here.

Chapter 5

Further Applications and Conclusions

The applications of the framework are not limited to the passenger weight data. Many data sets have subpopulation structure in them. We shortly demonstrate this with the delay data from the aviation industry. Further application and trends to use multivariate data and domain expertise are considered.

5.1 Weather-related Delays and on Applying the Framework

We shortly examine another application from the aviation industry to address a few issues that were not strongly present with our previous case example and enlighten further applications of the framework.

In passenger aviation understanding and predicting the flight delays is a significant matter for customer satisfaction and operating the flights altogether. The flight delay distributions happen to have similar a recognizable subpopulation structure, when we consider subpopulations based on destinations, flight operators, dates, times, weather conditions or other variables related to the flights.

In Figure 5.1 we look at the delays for each month. The shape is similar but there are more delays during the winter months. One of the reasons for this is the more unfavorable weather conditions during the winter mainly due the practicalities that are needed on the flight landing site.

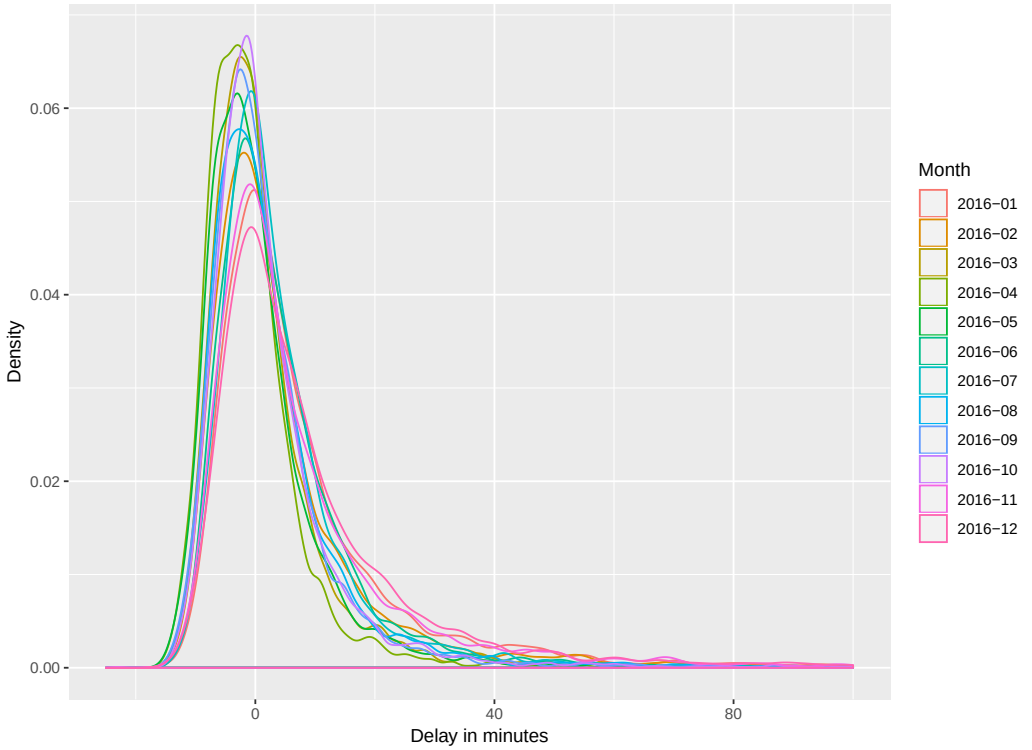


Figure 5.1: The flight delay distributions for each month.

An interesting operational question is to understand how the weather causes the delays for the flights. The framework may help to understand the amount of flights that arrive with certain delay interval and improve the accuracy to estimate the delays caused by extreme weather for which we have only a relatively few samples. The distribution structure for the subpopulations of different weather groups look similar to the monthly delays but the deviations and averages are greater on top of the fact that utilizing different distributions can be done much more rapidly than once a month when we have predictions of the weather conditions in the future.

This task is more complex than the previous passenger weight case and the clustering and variable selection methods presented may be inadequate. The weather data is very versatile and its relation to flights is very complex and

often bound to the area of operations, meaning that different airlines may need distinct models and data. Domain knowledge and appropriate feature selection are needed. As an example, instead of using every weather variable available, a suitable subset could be the METAR data that is a concise weather report for pilots. [10].

Similar problems have been tackled in aviation previously too. Herrema et al. [19] have developed a machine learning model for predicting abnormal runaway occupancy times in airway operations based on flight arrival data. The model identifies that many of the weather related variables are the most important, which are used in the actual model. Moreover, the model by Herrema et al. also essentially fits distributions to clustered groups like our framework. Though, the shape of distributions is chosen by just trying different families of distributions and subpopulation thinking is not otherwise present.

For this case of weather-related delays, in practice, ground operations could adjusted or flights could be canceled based on the statistical predictions by the framework. Furthermore, the successful clustering can help to us understand the delay risks or abnormalities [26], that are part of proactive flight managing that is taking place in the aviation industry, which different weather conditions also cause. However, to model the delays for each hours or even shorter time span we should expand the framework by including dependencies between delays that are close to each other. Figure 5.2 shows delays over one day. There seems to be some correlation, beyond what is caused just by the weather being the same, for flight delays that are close in time. Additional modeling, such a traffic congestion model [42] with using the time series data is needed to expand the framework to make operational prediction about the delays that would provide exact predictions instead of just general insights about the delays.

All in all, the framework provides interesting insights from the data just by on its own, if an adept analyst is there to construct the parts of the framework and interpret the outcomes. The framework is potentially applicable to any multivariate data and especially useful when rare incidents are considered, whether it is to understand passenger aviation or any other objective.

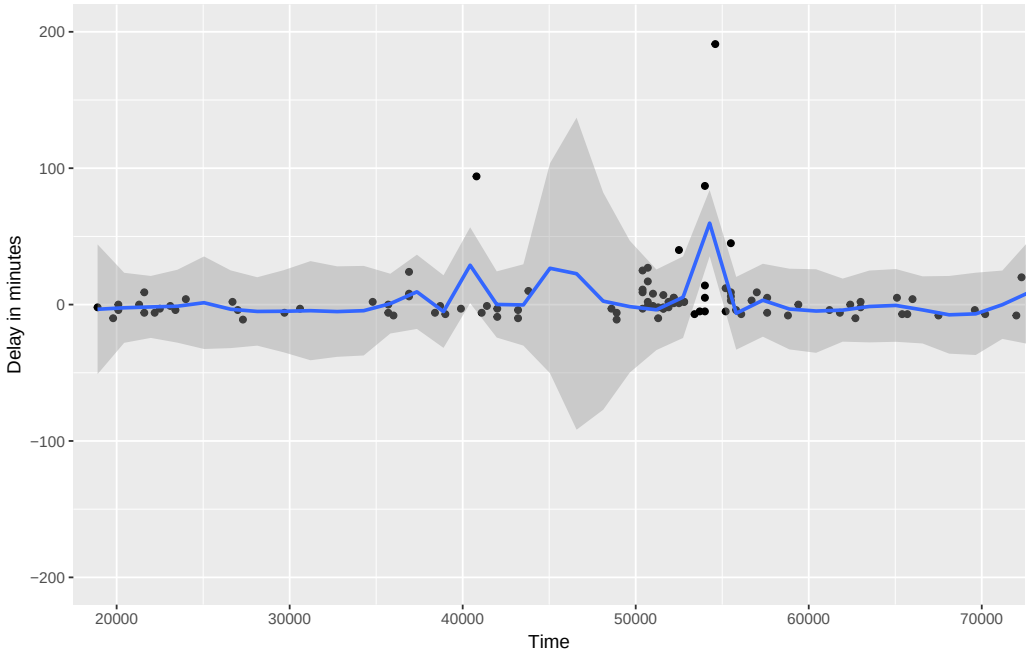


Figure 5.2: The flight delay time series for one day that shows congestion related patterns that may not be predicted, if the flight delay data points are assumed to be independent.

5.2 Conclusions

We introduced a framework to model the underlying subpopulation distribution structure a multivariate data might have. The framework used the entire data to cluster the data in a meaningful way so that the distributions for the variable of interest had understandable structure. The framework also required distribution fitting methods to actual construct the distributions from the samples. The core of understanding the common distribution structure was to identify the common shape for subpopulations. We used distribution statistics to determine the common shape and made improved estimates for the distributions of interest and their statistics. Our most defined method was the location-scale averaging method, that assumed that the shape could be understood by normalized versions of the distributions. We also discussed how higher moments and other statistics could be used for shape similarity identification and how the subpopulations could be weighted unevenly based on the statistical analysis or domain information by an expert.

The passenger weight case demonstrated how clustering methods of the framework found clear subpopulation structures from the multivariate survey data associated with the passenger weights. The clustering and their statistics were used to segment the network into meaningful groups for which different standard passenger weights could be operationally advantageous. The later part of this case considered how the framework could improve the accuracy of subpopulation distribution and consequently statistics estimates, such as extreme quantiles. The estimates by the method using the framework were shown to be more truthful than the trivial estimates. We also discussed further applications of the framework in the passenger aviation by taking a short look at flight delays and their relations to weather conditions.

The framework managed to provide novel insights about the data by exploring the subpopulation structures and giving more accurate estimates. The case examples showed how it worked in practice for passenger aviation data sets and operational interests. It should be noted that the framework needs to be adjusted based on the data and applications. The submethods, that is, the clustering, the distribution fitting and the common shape identification can be done in various ways, and the approaches shown in this thesis were just examples that were reasonable for the aviation data sets of interest. Especially, the theory and methods for the common shape identification should be developed, when the location-scale assumption does not hold very well. Other main issue that was not really present with our cases is that the subpopulation structure might be hard to find or it might not even exist. Here we found the structure pretty effortlessly which was due to the high quality of the data sets to understand the variable of interest.

Overall, the framework seems very prominent for understanding multivariate data sets in nature and business using the distributional subpopulation structure within them, but it needs both suitable submethods and data assets to attain truly usable insights and predictions.

Bibliography

[1] A. Azzalini and N. Torelli, Clustering via nonparametric density estimation, Statistics and Computing, 17 (2007), pp. 71–80.

[2] J. D. Banfield and A. E. Raftery, Model-based gaussian and non-gaussian clustering, Biometrics, (1993), pp. 803–821.

[3] Z. Berdowski, J. van den Broek-Serie, Y. K. Jetten, and Y. Kawabata, Survey on standard weights of passengers and baggage, European Aviation Safety Agency, Cologne, Germany, (2009).

[4] G. E. Box and D. R. Cox, An analysis of transformations, Journal of the Royal Statistical Society. Series B (Methodological), (1964), pp. 211–252.

[5] D. R. Cox, Principles of statistical inference, Cambridge university press, 2006.

[6] D. R. Cox and D. V. Hinkley, Theoretical statistics, Chapman and Hall/CRC, 1979.

[7] A. L. Dekkers, J. H. Einmahl, L. De Haan, et al., A moment estimator for the index of an extreme-value distribution, The Annals of Statistics, 17 (1989), pp. 1833–1855.

[8] M. L. Delignette-Muller, C. Dutang, et al., fitdistrplus: An R package for fitting distributions, Journal of Statistical Software, 64 (2015), pp. 1–34.

[9] European Union Aviation Safety Agency, Regulation (EU) 965/2012 on air operations, Annex IV – Part-CAT, EASA, (2018).

[10] Finnish Meteorological Institute, Lentosääpalvelut Suomessa, Guide book, (2013).

[11] C. Fraley and A. E. Raftery, Model-based clustering, discriminant analysis, and density estimation, Journal of the American statistical Association, 97 (2002), pp. 611–631.

[12] C. Fraley, A. E. Raftery, T. B. Murphy, and L. Scrucca, mclust version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation, tech. rep., Technical report, 2012.

[13] M. Ghosh, J. Rao, et al., Small area estimation: an appraisal, Statistical science, 9 (1994), pp. 55–76.

[14] J. C. Gower, A general coefficient of similarity and some of its properties, Biometrics, (1971), pp. 857–871.

[15] E. J. Gumbel, Statistical theory of extreme values and some practical applications, NBS Applied Mathematics Series, 33 (1954).

[16] I. Guyon and A. Elisseeff, An introduction to variable and feature selection, Journal of machine learning research, 3 (2003), pp. 1157–1182.

[17] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques, Elsevier, 2011.

[18] K. Haven, A. Majda, and R. Abramov, Quantifying predictability through information theory: small sample estimation in a non-gaussian framework, Journal of Computational Physics, 206 (2005), pp. 334–362.

[19] F. Herrema, V. Treve, B. Desart, R. Curran, and D. Visser, A novel machine learning model to predict abnormal runway occupancy times and observe related precursors, in 12th USA/Europe Air Traffic Management Research and Development Seminar, 2017.

[20] B. M. Hill, A simple general approach to inference about the tail of a distribution, The annals of statistics, (1975), pp. 1163–1174.

[21] J. R. M. Hosking, L-moments: Analysis and estimation of distributions using linear combinations of order statistics, Journal of the Royal Statistical Society: Series B (Methodological), 52 (1990), pp. 105–124.

[22] V. John, I. Angelov, A. Öncül, and D. Thévenin, Techniques for the reconstruction of a distribution from a finite number of its moments, Chemical Engineering Science, 62 (2007), pp. 2890–2904.

[23] M. Kuhn et al., Building predictive models in R using the caret package, Journal of statistical software, 28 (2008), pp. 1–26.

[24] S. X. Lee and G. J. McLachlan, Model-based clustering and classification with non-normal mixture distributions, Statistical Methods & Applications, 22 (2013), pp. 427–454.

[25] E. L. Lehmann and G. Casella, Theory of point estimation, Springer Science & Business Media, 2006.

[26] L. Li, S. Das, R. John Hansman, R. Palacios, and A. N. Srivastava, Analysis of flight data using clustering techniques for detecting abnormal operations, Journal of Aerospace information systems, 12 (2015), pp. 587–598.

[27] E. Limpert and W. A. Stahel, Problems with using the normal distribution – and ways to improve quality and efficiency of data analysis, PLoS One, 6 (2011), p. e21403.

[28] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton, Big data: the management revolution, Harvard business review, 90 (2012), pp. 60–68.

[29] G. J. McLachlan and K. E. Basford, Mixture models: inference and applications to clustering, vol. 84, M. Dekker New York, 1988.

[30] G. J. McLachlan, S. X. Lee, and S. I. Rathnayake, Finite mixture models, Annual review of statistics and its application, 6 (2019), pp. 355–378.

[31] D. Olson, Data mining in business services, Service Business, 1 (2007), pp. 181–193.

[32] D. Opitz and R. Maclin, Popular ensemble methods: an empirical study, Journal of artificial intelligence research, 11 (1999), pp. 169–198.

[33] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, Shape distributions, ACM Transactions on Graphics (TOG), 21 (2002), pp. 807–832.

[34] M. C. Peel, Q. Wang, R. M. Vogel, and T. A. McMAHON, The utility of L-moment ratio diagrams for selecting a regional probability distribution, Hydrological Sciences Journal, 46 (2001), pp. 147–155.

[35] D. Pfeffermann et al., New important developments in small area estimation, Statistical Science, 28 (2013), pp. 40–68.

[36] F. Provost and T. Fawcett, Data science and its relationship to big data and data-driven decision making, Big data, 1 (2013), pp. 51–59.

[37] J. N. Rao, Small-area estimation, Wiley StatsRef: Statistics Reference Online, (2014), pp. 1–8.

[38] R. A. Rigby and D. M. Stasinopoulos, Generalized additive models for location, scale and shape, Journal of the Royal Statistical Society: Series C (Applied Statistics), 54 (2005), pp. 507–554.

[39] Y. Rubner, C. Tomasi, and L. J. Guibas, The earth mover's distance as a metric for image retrieval, International journal of computer vision, 40 (2000), pp. 99–121.

[40] B. W. Silverman, Density estimation for statistics and data analysis, Routledge, 2018.

[41] A. Spanos, Probability theory and statistical inference: econometric modeling with observational data, Cambridge University Press, 1999.

[42] Y. Tu, M. O. Ball, and W. S. Jank, Estimating flight departure delay distributions? A statistical approach with long-term trend and short-term pattern, Journal of the American Statistical Association, 103 (2008), pp. 112–125.

[43] E. Vargo, R. Pasupathy, and L. Leemis, Moment-ratio diagrams for univariate distributions, Journal of Quality Technology, 42 (2010), pp. 276–286.

[44] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, Probability and statistics for engineers and scientists, vol. 5, Macmillan New York, 1993.

[45] A. E. Wegner, L. Ospina-Forero, R. E. Gaunt, C. M. Deane, and G. Reinert, Identifying networks with common organizational principles, Journal of Complex Networks, 6 (2018), pp. 887–913.

[46] A. Winkelbauer, Moments and absolute moments of the normal distribution, arXiv preprint arXiv:1209.4340, (2012).